

Dylan D. Schmorrow
Cali M. Fidopiastis (Eds.)

LNAI 10915

Augmented Cognition

Intelligent Technologies

12th International Conference, AC 2018
Held as Part of HCI International 2018
Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part I

1
Part I



 Springer

Lecture Notes in Artificial Intelligence

10915

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel

University of Alberta, Edmonton, Canada

Yuzuru Tanaka

Hokkaido University, Sapporo, Japan

Wolfgang Wahlster

DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann

DFKI and Saarland University, Saarbrücken, Germany

More information about this series at <http://www.springer.com/series/1244>

Dylan D. Schmorrow · Cali M. Fidopiastis (Eds.)

Augmented Cognition

Intelligent Technologies

12th International Conference, AC 2018
Held as Part of HCI International 2018
Las Vegas, NV, USA, July 15–20, 2018
Proceedings, Part I

Editors

Dylan D. Schmorow
Office of Naval Research
Orlando, FL
USA

Cali M. Fidopiastis
Design Interactive, Inc.
Orlando, FL
USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-91469-5 ISBN 978-3-319-91470-1 (eBook)
<https://doi.org/10.1007/978-3-319-91470-1>

Library of Congress Control Number: 2018943440

LNCS Sublibrary: SL7 – Artificial Intelligence

© Springer International Publishing AG, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

The 20th International Conference on Human-Computer Interaction, HCI International 2018, was held in Las Vegas, NV, USA, during July 15–20, 2018. The event incorporated the 14 conferences/thematic areas listed on the following page.

A total of 4,373 individuals from academia, research institutes, industry, and governmental agencies from 76 countries submitted contributions, and 1,170 papers and 195 posters have been included in the proceedings. These contributions address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The contributions thoroughly cover the entire field of human-computer interaction, addressing major advances in knowledge and effective use of computers in a variety of application areas. The volumes constituting the full set of the conference proceedings are listed in the following pages.

I would like to thank the program board chairs and the members of the program boards of all thematic areas and affiliated conferences for their contribution to the highest scientific quality and the overall success of the HCI International 2018 conference.

This conference would not have been possible without the continuous and unwavering support and advice of the founder, Conference General Chair Emeritus and Conference Scientific Advisor Prof. Gavriel Salvendy. For his outstanding efforts, I would like to express my appreciation to the communications chair and editor of *HCI International News*, Dr. Abbas Moallem.

July 2018

Constantine Stephanidis

HCI International 2018 Thematic Areas and Affiliated Conferences

Thematic areas:

- Human-Computer Interaction (HCI 2018)
- Human Interface and the Management of Information (HIMI 2018)

Affiliated conferences:

- 15th International Conference on Engineering Psychology and Cognitive Ergonomics (EPCE 2018)
- 12th International Conference on Universal Access in Human-Computer Interaction (UAHCI 2018)
- 10th International Conference on Virtual, Augmented, and Mixed Reality (VAMR 2018)
- 10th International Conference on Cross-Cultural Design (CCD 2018)
- 10th International Conference on Social Computing and Social Media (SCSM 2018)
- 12th International Conference on Augmented Cognition (AC 2018)
- 9th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management (DHM 2018)
- 7th International Conference on Design, User Experience, and Usability (DUXU 2018)
- 6th International Conference on Distributed, Ambient, and Pervasive Interactions (DAPI 2018)
- 5th International Conference on HCI in Business, Government, and Organizations (HCIBGO)
- 5th International Conference on Learning and Collaboration Technologies (LCT 2018)
- 4th International Conference on Human Aspects of IT for the Aged Population (ITAP 2018)

Conference Proceedings Volumes Full List

1. LNCS 10901, Human-Computer Interaction: Theories, Methods, and Human Issues (Part I), edited by Masaaki Kurosu
2. LNCS 10902, Human-Computer Interaction: Interaction in Context (Part II), edited by Masaaki Kurosu
3. LNCS 10903, Human-Computer Interaction: Interaction Technologies (Part III), edited by Masaaki Kurosu
4. LNCS 10904, Human Interface and the Management of Information: Interaction, Visualization, and Analytics (Part I), edited by Sakae Yamamoto and Hirohiko Mori
5. LNCS 10905, Human Interface and the Management of Information: Information in Applications and Services (Part II), edited by Sakae Yamamoto and Hirohiko Mori
6. LNAI 10906, Engineering Psychology and Cognitive Ergonomics, edited by Don Harris
7. LNCS 10907, Universal Access in Human-Computer Interaction: Methods, Technologies, and Users (Part I), edited by Margherita Antona and Constantine Stephanidis
8. LNCS 10908, Universal Access in Human-Computer Interaction: Virtual, Augmented, and Intelligent Environments (Part II), edited by Margherita Antona and Constantine Stephanidis
9. LNCS 10909, Virtual, Augmented and Mixed Reality: Interaction, Navigation, Visualization, Embodiment, and Simulation (Part I), edited by Jessie Y. C. Chen and Gino Fragomeni
10. LNCS 10910, Virtual, Augmented and Mixed Reality: Applications in Health, Cultural Heritage, and Industry (Part II), edited by Jessie Y. C. Chen and Gino Fragomeni
11. LNCS 10911, Cross-Cultural Design: Methods, Tools, and Users (Part I), edited by Pei-Luen Patrick Rau
12. LNCS 10912, Cross-Cultural Design: Applications in Cultural Heritage, Creativity, and Social Development (Part II), edited by Pei-Luen Patrick Rau
13. LNCS 10913, Social Computing and Social Media: User Experience and Behavior (Part I), edited by Gabriele Meiselwitz
14. LNCS 10914, Social Computing and Social Media: Technologies and Analytics (Part II), edited by Gabriele Meiselwitz
15. LNAI 10915, Augmented Cognition: Intelligent Technologies (Part I), edited by Dylan D. Schmorow and Cali M. Fidopiastis
16. LNAI 10916, Augmented Cognition: Users and Contexts (Part II), edited by Dylan D. Schmorow and Cali M. Fidopiastis
17. LNCS 10917, Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management, edited by Vincent G. Duffy
18. LNCS 10918, Design, User Experience, and Usability: Theory and Practice (Part I), edited by Aaron Marcus and Wentao Wang

19. LNCS 10919, Design, User Experience, and Usability: Designing Interactions (Part II), edited by Aaron Marcus and Wentao Wang
20. LNCS 10920, Design, User Experience, and Usability: Users, Contexts, and Case Studies (Part III), edited by Aaron Marcus and Wentao Wang
21. LNCS 10921, Distributed, Ambient, and Pervasive Interactions: Understanding Humans (Part I), edited by Norbert Streitz and Shin'ichi Konomi
22. LNCS 10922, Distributed, Ambient, and Pervasive Interactions: Technologies and Contexts (Part II), edited by Norbert Streitz and Shin'ichi Konomi
23. LNCS 10923, HCI in Business, Government, and Organizations, edited by Fiona Fui-Hoon Nah and Bo Sophia Xiao
24. LNCS 10924, Learning and Collaboration Technologies: Design, Development and Technological Innovation (Part I), edited by Panayiotis Zaphiris and Andri Ioannou
25. LNCS 10925, Learning and Collaboration Technologies: Learning and Teaching (Part II), edited by Panayiotis Zaphiris and Andri Ioannou
26. LNCS 10926, Human Aspects of IT for the Aged Population: Acceptance, Communication, and Participation (Part I), edited by Jia Zhou and Gavriel Salvendy
27. LNCS 10927, Human Aspects of IT for the Aged Population: Applications in Health, Assistance, and Entertainment (Part II), edited by Jia Zhou and Gavriel Salvendy
28. CCIS 850, HCI International 2018 Posters Extended Abstracts (Part I), edited by Constantine Stephanidis
29. CCIS 851, HCI International 2018 Posters Extended Abstracts (Part II), edited by Constantine Stephanidis
30. CCIS 852, HCI International 2018 Posters Extended Abstracts (Part III), edited by Constantine Stephanidis

<http://2018.hci.international/proceedings>



12th International Conference on Augmented Cognition

**Program Board Chair(s): Dylan D. Schmorrow
and Cali M. Fidopiastis, USA**

- Micah Clark, USA
- Martha Crosby, USA
- Dan Dolgin, USA
- Sven Fuchs, Germany
- Rodolphe Gentili, USA
- Scott Grigsby, USA
- Monte Hancock, USA
- Frank Hannigan, USA
- Robert Hubal, USA
- Øyvind Jøsok, Norway
- Ion Juvina, USA
- Benjamin Knott, USA
- Benjamin J. Knox, Norway
- Julie Marble, USA
- Chang S. Nam, USA
- Banu Onaral, USA
- Robinson Pino, USA
- Mannes Poel, The Netherlands
- Lauren Reinerman-Jones, USA
- Stefan Sütterlin, Norway
- Robert Sottolare, USA
- Ayoung Suh, Hong Kong, SAR China
- Christian Wagner, Hong Kong,
SAR China
- Melissa Walwanis, USA
- Quan Wang, USA
- Martin Westhoven, Germany

The full list with the Program Board Chairs and the members of the Program Boards of all thematic areas and affiliated conferences is available online at:

<http://www.hci.international/board-members-2018.php>



HCI International 2019

The 21st International Conference on Human-Computer Interaction, HCI International 2019, will be held jointly with the affiliated conferences in Orlando, FL, USA, at Walt Disney World Swan and Dolphin Resort, July 26–31, 2019. It will cover a broad spectrum of themes related to Human-Computer Interaction, including theoretical issues, methods, tools, processes, and case studies in HCI design, as well as novel interaction techniques, interfaces, and applications. The proceedings will be published by Springer. More information will be available on the conference website: <http://2019.hci.international/>.

General Chair

Prof. Constantine Stephanidis

University of Crete and ICS-FORTH

Heraklion, Crete, Greece

E-mail: general_chair@hcii2019.org

<http://2019.hci.international/>



Contents – Part I

Context Aware Adaptation Strategies in Augmented Cognition

Session Overview: Adaptation Strategies and Adaptation Management	3
<i>Sven Fuchs</i>	
Behaviour Adaptation Using Interaction Patterns with Augmented Reality Elements	9
<i>Marcel C. A. Baltzer, Christian Lassen, Daniel López, and Frank Flemisch</i>	
Adaptive, Policy-Driven, After Action Review in the Generalized Intelligent Framework for Tutoring	24
<i>Keith Brawner, Alan Carlin, Evan Oster, Chris Nucci, and Diane Kramer</i>	
Toward Adaptive Training Based on Bio-behavioral Monitoring	34
<i>Alexis Fortin-Côté, Daniel Lafond, Maëlle Kopf, Jean-François Gagnon, and Sébastien Tremblay</i>	
The Motivational Assessment Tool (MAT) Development and Validation Study	46
<i>Elizabeth Lameier, Lauren Reinerman-Jones, Gerald Matthews, Elizabeth Biddle, and Michael Boyce</i>	
A Multi-sensor Approach to Linking Behavior to Job Performance	59
<i>Alison M. Perez, Amanda E. Kraft, Raquel Galvan-Garza, Matthew Pava, Amanda Barkan, William D. Casebeer, and Matthias D. Ziegler</i>	
Leveraging Cognitive Psychology Principles to Enhance Adaptive Instruction	69
<i>Anne M. Sinatra</i>	
Community Models to Enhance Adaptive Instruction	78
<i>Robert Sottolare</i>	
Biocybernetic Adaptation Strategies: Machine Awareness of Human Engagement for Improved Operational Performance	89
<i>Chad Stephens, Frédéric Dehais, Raphaëlle N. Roy, Angela Harrivel, Mary Carolyn Last, Kellie Kennedy, and Alan Pope</i>	

Brain Sensors and Measures for Operational Environments

Do Not Disturb: Psychophysiological Correlates of Boredom, Flow and Frustration During VR Gaming.	101
<i>Klaas Bombeke, Aranka Van Dongen, Wouter Durnez, Alessandra Anzolin, Hannes Almgren, Anissa All, Jan Van Looy, Lieven De Marez, Daniele Marinazzo, and Elena Patricia Núñez Castellar</i>	
M.I.N.D. Brain Sensor Caps: Coupling Precise Brain Imaging to Virtual Reality Head-Mounted Displays	120
<i>Gyoung Kim, Joonhyun Jeon, and Frank Biocca</i>	
Assessing Operator Psychological States and Performance in UAS Operations	131
<i>Jinchao Lin, Gerald Matthews, Lauren Reinerman-Jones, and Ryan Wohleber</i>	
Trust in Sensing Technologies and Human Wingmen: Analogies for Human-Machine Teams	148
<i>Joseph B. Lyons, Nhut T. Ho, Lauren C. Hoffmann, Garrett G. Sadler, Anna Lee Van Abel, and Mark Wilkins</i>	
Deep Convolutional Neural Networks and Power Spectral Density Features for Motor Imagery Classification of EEG Signals	158
<i>A. F. Pérez-Zapata, A. F. Cardona-Escobar, J. A. Jaramillo-Garzón, and Gloria M. Díaz</i>	
Long Term Use Effects of a P300-Based Spelling Application	170
<i>Cristian-Cezar Postelnicu, Florin Girbacia, Octavian Machidon, and Gheorghe-Daniel Voinea</i>	
A Wearable Multisensory, Multiagent Approach for Detection and Mitigation of Acute Cognitive Strain: Phase I - Vocalization analysis	180
<i>Anil Raj, Brooke Roberts, Kristy Hollingshead, Neil McDonald, Melissa Poquette, and Walid Soussou</i>	
Classification Procedure for Motor Imagery EEG Data	201
<i>Ellton Sales Barros and Nelson Neto</i>	
WebBCI: An Electroencephalography Toolkit Built on Modern Web Technologies.	212
<i>Pierce Stegman, Chris Crawford, and Jeff Gray</i>	
A Cross-Brain Interaction Platform Based on Neurofeedback Using Electroencephalogram.	222
<i>Rongrong Zhang and Xiaojie Zhao</i>	

Single-Channel EEG Sleep Stage Classification Based
on K-SVD Algorithm 231
Shigang Zuo and Xiaojie Zhao

Artificial Intelligence and Machine Learning in Augmented Cognition

Improving Automation Transparency: Addressing Some of Machine
Learning’s Unique Challenges 245
Corey K. Fallon and Leslie M. Blaha

Artificial Intelligence for Advanced Human-Machine Symbiosis 255
Scott S. Grigsby

Feature Extraction from Social Media Posts for Psychometric Typing
of Participants. 267
*Charles Li, Monte Hancock, Ben Bowles, Olivia Hancock, Lesley Perg,
Payton Brown, Asher Burrell, Gianella Frank, Frankie Stiers,
Shana Marshall, Gale Mercado, Alexis-Walid Ahmed,
Phillip Beckelheimer, Samuel Williamson, and Rodney Wade*

Intermediate Information Grouping in Cluster Recognition 287
*Chloe Chun-wing Lo, Markus Hollander, Freda Wan,
Alexis-Walid Ahmed, Nikki Bernobić, Nick Nuon, and Michael Shrider*

Human-Machine Teaming and Cyberspace 299
Fernando J. Maymí and Robert Thomson

Automatically Unaware: Using Data Analytics to Detect
Physiological Markers of Cybercrime 316
Nancy Mogire, Randall K. Minas, and Martha E. Crosby

Understanding Behaviors in Different Domains: The Role of Machine
Learning Techniques and Network Science. 329
*Grace Teo, Lauren Reinerman-Jones, Joseph McDonnell,
Hayden J. Trainor, Rainier A. Porras, and Jacob G. Feuerman*

A Workflow for Network Analysis-Based Structure Discovery
in the Assessment Community 341
*Grace Teo, Lauren Reinerman-Jones, Mark E. Riecken,
Joseph McDonnell, Scott Gallant, Maartje Hidalgo,
and Clayton W. Burford*

Augmented Cognition in Virtual and Mixed Reality

Immersion Versus Embodiment: Embodied Cognition for Immersive
Analytics in Mixed Reality Environments 355
Denis Gračanin

Development and Application of the Hybrid Space App for Measuring Cognitive Focus in Hybrid Contexts 369
Øyvind Jøsok, Mathias Hedberg, Benjamin J. Knox, Kirsi Helkala, Stefan Sütterlin, and Ricardo G. Lugo

Identifying Affordance Features in Virtual Reality: How Do Virtual Reality Games Reinforce User Experience? 383
Jumin Lee, Jounghae Bang, and Hyunju Suh

Augmented Reality and Telestrated Surgical Support for Point of Injury Combat Casualty Care: A Feasibility Study 395
Geoffrey T. Miller, Tyler Harris, Y. Sammy Choi, Stephen M. DeLellis, Kenneth Nelson, and J. Harvey Magee

Cultivating Environmental Awareness: Modeling Air Quality Data via Augmented Reality Miniature Trees 406
Jane Prophet, Yong Ming Kow, and Mark Hurry

Enhancing Audience Engagement Through Immersive 360-Degree Videos: An Experimental Study 425
Ayoung Suh, Guan Wang, Wenying Gu, and Christian Wagner

Enhancing Bicycle Safety Through Immersive Experiences Using Virtual Reality Technologies 444
Hiroki Tsuboi, Shuma Toyama, and Tatsuo Nakajima

Author Index 457

Contents – Part II

Cognitive Modeling, Perception, Emotion and Interaction

Multi-modal Interruptions on Primary Task Performance	3
<i>Pooja P. Bovard, Kelly A. Sprehn, Meredith G. Cunha, Jaemin Chun, SeungJun Kim, Jana L. Schwartz, Sara K. Garver, and Anind K. Dey</i>	
Can University Students Use Basic Breathing Activities to Regulate Physiological Responses Caused by Computer Use? A Pilot Study	15
<i>Hubert K. Brumback</i>	
Human Performance Augmentation in Context: Using Artificial Intelligence to Deal with Variability—An Example from Narrative Influence.	32
<i>William D. Casebeer, Matthias Ziegler, Amanda E. Kraft, Jason Poleski, and Bartlett Russell</i>	
Human Machine Interactions: Velocity Considerations.	43
<i>Joseph Cottam, Leslie M. Blaha, Kris Cook, and Mark Whiting</i>	
Strengthening Health and Improving Emotional Defenses (SHIELD).	58
<i>Seth Elkin-Frankston, Arthur Wollocko, and James Niehaus</i>	
Assessment of Wearable Tactile System: Perception, Learning, and Recall . . .	67
<i>Linda R. Elliott, Bruce J. P. Mortimer, Rodger A. Pettitt, and Robert E. Wooldridge</i>	
Visualization of Network Security Data by Haptic.	78
<i>Manabu Ishihara and Taiki Kanayama</i>	
Using Scenarios to Validate Requirements Through the Use of Eye-Tracking in Prototyping.	94
<i>Tia Larsen-Calcano, Omar Ochoa, and Richard Simonson</i>	
Measuring Focused Attention Using Fixation Inner-Density	105
<i>Wen Liu, Soussan Djamasbi, Andrew C. Trapp, and Mina Shojaeizadeh</i>	
Cognition and Predictors of Password Selection and Usability	117
<i>Lila A. Loos and Martha E. Crosby</i>	
Forget the Password: Password Memory and Security Applications of Augmented Cognition	133
<i>Nancy Mogire, Michael-Brian Ogawa, Randall K. Minas, Brent Auernheimer, and Martha E. Crosby</i>	

Designing and Evaluating Reporting Systems in the Context of New Assessments	143
<i>Diego Zapata-Rivera, Priya Kannan, Carol Forsyth, Stephanie Peters, Andrew D. Bryant, Enruo Guo, and Rodolfo Long</i>	
Human Augmentation of UAV Cyber-Attack Detection	154
<i>Haibei Zhu, Mahmoud Elfar, Miroslav Pajic, Ziyao Wang, and Mary L. Cummings</i>	
Augmented Learning and Training	
Mitigating Skill Decay in Military Instruction and Enemy Analysis via GIFT	171
<i>Michael W. Boyce, Jeanine A. DeFalco, Robert C. Davis, Erik K. Kober, and Benjamin Goldberg</i>	
Developing Accelerated Learning Models in GIFT for Medical Military and Civilian Training.	183
<i>Jeanine A. DeFalco, R. Stanley Hum, and Michael Wilhelm</i>	
Experiential Intelligent Tutoring: Using the Environment to Contextualize the Didactic	192
<i>Benjamin Goldberg and Michael Boyce</i>	
Guided Mindfulness: Optimizing Experiential Learning of Complex Interpersonal Competencies	205
<i>Richard L. Griffith, Lisa A. Steelman, Nicholas Moon, Sherif al-Qallawi, and Nisha Quraishi</i>	
Curriculum for Accelerated Learning Through Mindfulness (CALM).	214
<i>Anna Skinner, Cali Fidopiastis, Sebastian Pascarelle, and Howard Reichel</i>	
Augmented Reality for Tactical Combat Casualty Care Training	227
<i>Glenn Taylor, Anthony Deschamps, Alyssa Tanaka, Denise Nicholson, Gerd Bruder, Gregory Welch, and Francisco Guido-Sanz</i>	
A Workload Comparison During Anatomical Training with a Physical or Virtual Model.	240
<i>Andrew Wismer, Lauren Reinerman-Jones, Grace Teo, Sasha Willis, Kelsey McCracken, and Matthew Hackett</i>	

Shared Cognition, Team Performance and Decision-Making

Parole Board Personality and Decision Making Using Bias-Based Reasoning	255
<i>Katy Hancock, Payton Brown, Antoinette Hadgis, Markus Hollander, and Michael Shrider</i>	
Validation of a Maritime Usability Study with Eye Tracking Data.	273
<i>Odd Sveinung Hareide and Runar Ostnes</i>	
The Wide Area Virtual Environment: A New Paradigm for Medical Team Training	293
<i>Alan Liu, Eric Acosta, Jamie Cope, Valerie Henry, Fernando Reyes, Joseph Bradascio, and Wesley Meek</i>	
Using Bots in Strategizing Group Compositions to Improve Decision-Making Processes	305
<i>Shai Neumann, Suraj Sood, Markus Hollander, Freda Wan, Alexis-Walid Ahmed, and Monte Hancock</i>	
Augmenting Clinical Performance in Combat Casualty Care: Telemedicine to Automation.	326
<i>Jeremy C. Pamplin, Ronald Yeaw, Gary R. Gilbert, Konrad L. Davis, Elizabeth Mann-Salinas, Jose Salinas, Daniel Kral, and Loretta Schlachta-Fairchild</i>	
Optimizing Team Performance When Resilience Falters: An Integrated Training Approach	339
<i>Debbie Patton, Lisa Townsend, Laura Milham, Joan Johnston, Dawn Riddle, Amanda R. Start, Amy B. Adler, and Karen Costello</i>	
A Human Perspective on Maritime Autonomy	350
<i>Tore Relling, Margareta Lützhöft, Runar Ostnes, and Hans Petter Hildre</i>	
Improving Understanding of Mindfulness Concepts and Test Methods.	363
<i>Melissa M. Walwanis and Derek S. Bryan</i>	
Author Index	375

Context Aware Adaptation Strategies in Augmented Cognition



Session Overview: Adaptation Strategies and Adaptation Management

Sven Fuchs^(✉)

Fraunhofer Institute for Communication, Information Processing
and Ergonomics FKIE, 53343 Wachtberg, Germany
sven.fuchs@fkie.fraunhofer.de

Abstract. Researchers in the field of Augmented Cognition (AugCog) have often focused on detecting and classifying cognitive problem states (e.g. via physiological sensors). Yet, knowing what cognitive state to address through adaptation is merely a first step to building an adaptive system. The next steps – to determine what to adapt and how to do it – are just as interesting and challenging. There is great untapped potential in the adaptation component, yet it has been underrepresented and underappreciated in the AugCog community discourse. The goal of this contribution and the associated conference session is to get our community thinking about how to put their cognitive state diagnoses to use in an innovative manner, how to develop innovative adaptation strategies, and how to address adaptation management issues. This session overview lists a number of challenges faced by AugCog researchers with respect to the development of adaptation strategies and adaptation management frameworks. The papers featured in this session encompass a diverse set of approaches and ideas to address these challenges.

Keywords: Augmented cognition · Adaptive systems · Adaptive training
Human-systems integration · Adaptive human-computer interaction
Cooperative systems · Adaptation strategies · Adaptation management
Adaptation frameworks

1 The Challenge of Effective Adaptation Management

The operator tried and tried but the task was too hard. Then suddenly, the task disappeared. The operator did not realize that physiological sensors in his Augmented Cognition (AugCog) system detected a critical state – high cognitive load – and in response triggered an automation strategy to reduce task load. What sounds like a reasonable approach – if the operator is overloaded, automate certain tasks to reduce the task load, and as a result cognitive load will decrease as well – may not lead to the desired result. What if task load was not even high and a lack of experience caused the operator to experience high cognitive load? In that case, automation could in fact be counterproductive, as it would no longer allow the user to gain the necessary experience.

In a different fictional AugCog system, oculomotor metrics are used to evaluate the attentional focus of the operator. How should the system react to an inappropriate focus?

Employ cueing strategies to shift the operator's focus? Declutter the display to minimize distraction? The effectiveness of an adaptation would greatly depend on whether the operator missed a task due to over-engagement and attentional tunneling [1], or if the user fell victim to vigilance decrement and task-related fatigue after a long shift.

These examples illustrate the ease of underestimating the challenge of adaptation management. Adaptation management involves selecting and configuring appropriate and effective adaptation strategies to address detected problem states, but also monitoring their effects and effectiveness. As AugCog diagnostics detect opportunities for adaptation, adaptation strategies are usually triggered in response to a specific diagnostic outcome. As situation and context evolve, and as the effects of the adaptation kick in, that specific situation is no longer present and a once adequate adaptation strategy may become inadequate. Continued adaptation may even have negative effects on the operator and task performance, as it may occupy cognitive resources, interrupt a high priority task, or otherwise affect the operator's attention inadequately. As an example, automating a task in a high-stress/high-workload phase may be helpful to maintain performance through this phase. Keep automating longer than necessary, however, and issues with automation complacency (e.g., [2]) and out-of-the-loop performance problems (e.g. [3, 4]) may arise.

Hence, the benefit of an adaptation should outweigh its potential cost. Cognitive costs of adaptation have been demonstrated in past implementations of adaptive human-machine interaction. For example, Dorneich et al. [5] report "a loss of situation awareness and survey knowledge of the environment" (p. iv), participants in another study reported confusion and impressions of inconsistency in the information display [6]. Fuchs et al. provide an overview of potential costs and benefits for a number of adaptation strategies previously used in Augmented Cognition systems [7]. The effects are similar to those observed in interactions with automated systems. Lessons learned from automation research (e.g. [8, 9]) should therefore be considered when designing adaptation strategies and adaptation management frameworks.

Adaptation management is a relevant topic for both adaptive operational environments and adaptive training systems; however, the objectives of adaptation are quite different between the two. In operational environments, the overall goal of adaptation is to optimize performance – adequate performance is the target state. Adaptive training aim to optimize training effectiveness and efficiency. To that end, good performance may present an opportunity to accelerate training and therefore indicate a need for adaptation. The same is true for critical cognitive states: in operational environments, degraded cognition poses safety risks and should be addressed. A training system may intentionally induce such states to train self-regulation or coping strategies. Finally, while an operational adaptive system would likely intervene to avoid human error, errors are acceptable and even desirable in training, as they offer learning opportunities.

As briefly outlined above, the challenges associated with adaptation management are substantial and remain largely unaddressed. This session on "Adaptation Strategies and Adaptation Management" aims to raise awareness for this essential component of AugCog systems and initiate scientific discourse to address these issues in future research.

2 Session Themes

The session opens with a look back at 15 years of AugCog research. Dylan Schmorow, who founded and led the field of Augmented Cognition while serving as a Program Manager at the Defense Advanced Research Projects Agency (DARPA), will review past efforts, provide lessons-learned relevant for adaptation management, and share his vision for a 21st century human-computer symbiosis. Further contributions address more specific challenges but can be categorized into five broader themes discussed in the following sections.

2.1 Enhancing Adaptation Through Context

Some AugCog systems have relied solely on physiological indicators of cognitive states to inform adaptation. However, it has been claimed that effective adaptation management requires contextual data as it contains crucial information about the state of the system, the task, and the user. Without such information, adaptations may be triggered or withdrawn at inopportune moments, potentially disrupting or confusing the user, or leading to task switching issues, situation awareness problems, and workload increases. In these cases, adaptations may even have a negative impact on performance, outweighing the benefit of adaptation altogether. This inconsiderate use of adaptation strategies has been labeled “brute force mitigation” [10]. In contrast, context-aware adaptation frameworks would process not only information about the physiological state of the user, but also behavioral data, task state, environmental parameters, sensor information, system events, or user interactions and preferences [11]. These can then be interpreted to derive task context, possible root causes for observed performance problems, or user intent to dynamically select and configure appropriate adaptations at runtime (cf. [12]).

In this session, Baltzer et al. [13] provide insights into a context-sensitive adaptation mechanism for cooperative guidance and control of highly automated vehicles. Their conceptual approach of analyzing “interaction patterns” (patterns that combine driver activity and environmental parameters) is used to adapt driver assistance systems to the situation at hand and determine whether and how technical intervention is necessary.

2.2 Adaptation Management in Adaptive Training Environments

To increase training efficiency, many adaptive training systems detect when a trainee is ready to move on. One challenge in this domain is to find appropriate indicators to advance training to the next level.

In this session, Stephens et al. [14] provide an overview of mental states of interest for adaptive training and various approaches to operationalize them. Fortin-Côte et al. [15] report a study to determine an optimal trigger rule for adapting an adaptive training environment based on different combinations of workload and performance metrics.

2.3 Stages of Adaptation

Consider a state of high cognitive load that is addressed through an automation strategy. With automation active, tasks are offloaded from the user and cognitive workload may decrease to an uncritical level. Subsequent withdrawal of the automation, however, would lead to an increase in workload that, again, triggers adaptation. To avoid rapid adaptive state oscillation and system instability, it may thus not be sufficient to merely switch adaptation strategies on and off. One approach to avoiding this instability is a gradual approach to adaptation that is as restrained as possible but as intrusive as necessary. Tollar [16] suggests adaptations that are “graduated or delivered incrementally in levels to help operators to keep ‘in the groove’.” (p. 417). Fuchs et al. describe an approach that uses “Stages of Adaptation” to modulate the intensity and intrusiveness of adaptation based on task priority [17].

In this session, Baltzer et al. [13] present a “stepwise escalation” approach to provide context-adequate intervention in a cooperative driving task. Based on context information, an automated vehicle will dynamically add auditory and/or tactile cues to communicate a detected obstacle. If deemed necessary, the vehicle will intervene and decouple the driver from the task to initiate an appropriate maneuver.

2.4 The Adaptive Operator

Humans themselves are adaptive systems (cf. [18, 19]). They react and adapt to changing task demands within a “zone of adaptability” [20]. These adaptations may be voluntary or involuntary. For example, humans may consciously decide to invest more effort to perform better if deemed necessary, or stress may cause the release of hormones leading to higher arousal and alertness. In the context of adaptive automation, Veltman and Jansen [19] expect that adaptive technical systems are more likely to work successfully if they started reallocating tasks only when the operator’s intrinsic adaptation mechanisms are no longer able to adequately react to changing task demands. Otherwise, two adaptive systems (the “adaptive operator” and the adaptive technical system) may interact in a counterproductive manner.

Stephens et al. [14] contribute an interesting twist to the “adaptive operator” theme as they present an overview of adaptive systems aimed at enhancing the operator’s self-awareness and self-regulation. Instead of adapting the system to achieve optimal performance or efficiency, the system provides feedback to improve self-monitoring and self-regulation skills that will help the user maintain more effective mental states under critical conditions.

2.5 Machine Learning Approaches for Adaptation Management

Machine learning approaches have been extensively used for cognitive state classification, but they may also prove useful for adaptation management. Evaluating the success of certain strategies and adaptation mechanisms in real-time, learning user characteristics, strategies, and preferences, and understanding and reacting to the effects of changing conditions are all aspects that could benefit from machine learning and artificial intelligence.

Adaptive training approaches have used machine learning to identify learner needs and tailor the learning experience to the individual learner. For machine learning algorithms to be effective, however, large amounts of individual training data are necessary. In this session, Sottolare [21] presents an idea to overcome this major limitation. He proposes to employ the concept of personas to develop a community-based learner model that represents a typical learner and evolves over time as the community provides additional data points.

3 The Road Ahead

The challenges of adaptation management are manifold due to the highly dynamic nature of human operators, their experience, their strategies, and the complex task environments. This session overview provided a number of adaptation-related themes that will hopefully spark interest in the community and inspire future research.

As Augmented Cognition systems move into the real world, it is time to embrace the true level of complexity of these highly integrated human-machine systems. Outside the laboratory, it will no longer be sufficient to observe problem state X and trigger adaptation strategy A_X in response. Future adaptation managers should be flexible enough to detect and account for unanticipated system states and correct or expand their future expectations and reactions as necessary. Effective AugCog systems will process extensive amounts of contextual data and use holistic models of cognition that consider the interplay and interdependencies of multiple cognitive states (cf. [22]) to dynamically mitigate the true source of detected problems.

References

1. Wickens, C.D.: Attentional Tunneling and Task Management. Technical report AHFD-05-01/ NASA-05-10. NASA Ames Research Center, Moffett Field, CA (2005)
2. Parasuraman, R., Molloy, R., Singh, I.L.: Performance consequences of automation-induced complacency. *Int. J. Aviat. Psychol.* **3**(1), 1–23 (1993)
3. Kessel, C.J., Wickens, C.D.: The transfer of failure-detection skills between monitoring and controlling dynamic systems. *Hum. Factors* **24**, 49–60 (1982)
4. Endsley, M.R., Kiris, E.O.: The out-of-the-loop performance problem and level of control in automation. *Hum. Factors* **37**(2), 381–394 (1995)
5. Dorneich, M., Whitlow, S., Ververs, P.M., Mathan, S., Raj, A., Muth, E., Hoover, A., DuRousseau, D., Parra, L., Sajda, P.: DARPA Improving Warfighter Information Intake under Stress - Augmented Cognition Concept Validation Experiment (CVE) Analysis Report for the Honeywell Team. DARPA/IPTO Technical report (2004)
6. Hale, K.S., Fuchs, S., Berka, C., Levendowski, D., Axelsson, P., Baskin, A., Juhnke, J.: Information Delivery and Display for Shared Awareness in the Net-Centric Battlespace. SBIR Phase I Final Technical report under Contract W31P4Q-06-C-0041. Design Interactive, Inc., Oviedo, FL (2006)
7. Fuchs, S., Hale, K.S., Stanney, K.M., Juhnke, J., Schmorow, D.D.: Enhancing mitigation in augmented cognition. *J. Cogn. Eng. Decis. Making* **3**, 309–326 (2007)

8. Breton, R., Bossé, É.: The cognitive costs and benefits of automation. In: *The Role of Humans in Intelligent and Automated Systems. Proceedings of the RTO Human Factors and Medicine Panel (HFM) Symposium (RTO-MP-088)*. NATO RTO, Neuilly-sur-Seine, France (2003)
9. Endsley, M.R.: Automation and situation awareness. In: Parasuraman, R., Mouloua, M. (eds.) *Automation and Human Performance: Theory and Applications*, pp. 163–181. Lawrence Erlbaum, Mahwah (1996)
10. Stanney, K., Reeves, L.: Mitigation strategies and performance effects. White paper outbrief from a working session at *Improving Warfighter Information Intake Under Stress*, AugCog PI Meeting, Chantilly, VA (2005)
11. Schwarz, J., Fuchs, S.: Multidimensional real-time assessment of user state and performance to trigger dynamic system adaptation. In: Schmorow, Dylan D., Fidopiastis, Cali M. (eds.) *AC 2017. LNCS (LNAI)*, vol. 10284, pp. 383–398. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58628-1_30
12. Fuchs, S., Schwarz, J.: Towards a dynamic selection and configuration of adaptation strategies in augmented cognition. In: Schmorow, Dylan D., Fidopiastis, Cali M. (eds.) *AC 2017. LNCS (LNAI)*, vol. 10285, pp. 101–115. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_7
13. Baltzer, C.A., Lassen, C., López, D., Flemisch, F.: Behaviour adaptation using interaction patterns with augmented reality elements. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) *HCI 2018. LNCS (LNAI)*, vol. 10915, pp. 9–21. Springer, Cham (2018)
14. Stephens, C., Dehais, F., Roy, R., Harnivel, A., Last, M. C., Kennedy, K., Pope, A.: Biocybernetic adaptation strategies: machine awareness of human state for improved operational performance. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) *HCI 2018. LNCS (LNAI)*, vol. 10915, pp. 89–98. Springer, Cham (2018)
15. Fortin-Côte, A., Lafond, D., Kopf, M., Gagnon, J.-F., Tremblay, S.: Adaptive training based on biobehavioral monitoring. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) *HCI 2018. LNCS (LNAI)*, vol. 10915, pp. 34–45. Springer, Cham (2018)
16. Tollar, J.T.: Statistical process control as a triggering mechanism for augmented cognition mitigations. In: Schmorow, D.D. (ed.) *Foundations of Augmented Cognition*, pp. 414–420. Lawrence Erlbaum Associates, Mahwah (2005)
17. Fuchs, S., Hale, K.S., Stanney, K.M., Berka, C., Levendowski, D., Juhnke, J.: Physiological sensors cannot effectively drive system mitigation alone. In: Schmorow, D.D., Stanney, K. M., Reeves, L.M. (eds.) *Foundations of Augmented Cognition*, 2nd edn, pp. 193–200. Strategic Analysis Inc., Arlington (2006)
18. Wiener, N.: *The Human Use of Human Beings*. Houghton Mifflin, Boston (1950)
19. Veltman, J.A., Jansen, C.: The adaptive operator. In: Vincenzi, D.A., Mouloua, M., Hancock, P. (eds.) *Human Performance, Situation Awareness, and Automation: Current Research and Trends*, vol. 2, pp. 7–10. Lawrence Erlbaum Associates, Mahwah (2004)
20. Hancock, P.A., Chignell, M.H.: Input information requirements for an adaptive human-machine system. In: *Proceedings of the Tenth Department of Defense Conference on Psychology*, pp. 493–498. Defense Technical Information Center, Colorado Springs (1986)
21. Sottillare, R. A.: Community Models to Enhance Adaptive Instruction. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) *HCI 2018. LNCS (LNAI)*, vol. 10915, pp. 78–88. Springer, Cham (2018)
22. Schwarz, J., Fuchs, S., Flemisch, F.: Towards a more holistic view on user state assessment in adaptive human-computer interaction. In: *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1247–1253. IEEE, San Diego (2014)



Behaviour Adaptation Using Interaction Patterns with Augmented Reality Elements

Marcel C. A. Baltzer¹✉, Christian Lassen¹, Daniel López¹,
and Frank Flemisch^{1,2}

¹ Fraunhofer FKIE, Fraunhoferstraße 20, 53343 Wachtberg, Germany
marcel.baltzer@fkie.fraunhofer.de

² RWTH Aachen University, Bergdriesch 27, 52062 Aachen, Germany

Abstract. This publication describes a systematic approach for behaviour adaptations of humans, based on interaction patterns as a fundamental way to design and describe human machine interaction, and on image schemas as the basic elements of the resulting interaction. The natural learning path since childhood involves getting knowledge by experience; it is during this process that image schemas are built. The approach described in this paper was developed in close interplay with the concepts of cooperative guidance and control (CGC), where a cooperative automation and a human control a machine together, and of augmented reality (AR), where a natural representation of the world, e.g. in form of a video stream, is enriched with dynamic symbology. The concept was instantiated as interaction patterns “longitudinal and lateral collision avoidance”, implemented in a fix based simulator, and tested with professional operators whether driving performance and safety in a vehicle with restricted vision could be improved. Furthermore, it was tested whether interaction patterns could be used to adapt the current driver behaviour towards better performance while reducing the task load. Using interaction patterns that escalated according to the drivers actions and the current environmental state, lead to a reduction of temporal demand, effort and frustration. Furthermore less collisions were counted and the overall lateral displacement of the vehicle was reduced. The results were a good mix of encouragement and lessons learned, both for the methodical approach of pattern based human machine interaction, and for the application of AR-based cooperative guidance and control.

Keywords: Augmented reality · Interaction patterns
Human behaviour adaptation · Intelligent transportation systems
Cooperative guidance and control

1 Introduction

Military vehicles like heavy trucks, tanks or excavators face the challenge of being sufficiently guarded against enemy fire and being safely manoeuvrable. The latter includes driving capability but also vision in order to see where the vehicle is driving. The safety aspect is not only important for the driver himself but also for other road users. Current military vehicles therefore face a trade-off between optimal vision and optimal armour resulting in more obstruction of the driver’s sight and the necessity of

an assisting co-driver. One option to overcome such limitations is to create a virtually transparent vehicle by using a camera-monitor system that provides a seamless vision to the driver.

Another important aspect of shielded vehicles in combat scenarios is situational awareness. Situational awareness (SA) is a decisive aspect of survivability for combat vehicles, and AR technologies have the potential to take SA to the next level. The introduction of optical, IR and acoustic sensors for monitoring the vehicle's surroundings, and in particular the integration of Battlefield Management Systems (BMS), has significantly improved situational awareness. However, in time critical situations the vehicle crew (e.g. commander, gunner and driver) have difficulties fully exploiting the information provided by these systems since the crew has to focus on the scene where the action is taking place. Augmentation offer the potential of overcoming this deficiency by using for example AR technologies displaying information, typically from the BMS, directly in the operator's sight in form of graphical symbols depending on the driver's current actions. Thus the operator can pay full attention to what is going on in the vehicle's vicinity while staying updated on the tactical situation within the sights' field of view.

The idea presented in this paper is to use interaction patterns to adapt assistance to the current situation formed by the environment and the driver's behaviour and actions. The behaviour of interaction patterns depends on the driver's actions and the state of the environment. Their focus is to adapt the driver's actions towards a safer system state through multimodal interaction.

First, the general concept of interaction patterns will be described. Second the application of interaction patterns in shielded vehicles will be presented and the respective outcomes of a conducted study. The paper finishes with a discussion of the findings and concludes with an outlook towards other domains.

2 Pattern Languages

The idea of using Pattern Languages has been discussed and applied for several years now. It was first introduced by Alexander in the area of architecture where he explains how a set of recurring designs can allow anyone to create its own house tailored to his needs by using previously known and tested solutions [1]. These solutions, so called Patterns, have the characteristic that they can be linked to one another allowing the user for complex and unique creations similarly to how a spoken language permits creating an unlimited quantity of sentences by arranging the words.

Pattern languages have found usages outside of architecture making them powerful tools that can help multidisciplinary teams communicate better without needing to have a complete understanding of how a process works while also helping to speed up the development of new tools. One of these areas is in the development of software where the pattern language approach can be used for object oriented programming. Here, the objects and classes are described in a more general way that can later be adapted to solve a specific problem [2]. Gamma further elaborates on the description and elements of a pattern expanding on the definition by Alexander. According to Gamma, a pattern has four essential elements: a name, a problem description, a solution and

consequences of utilizing the pattern. The basic structure was also the starting point for the patterns described in this paper.

In the area of Interaction Design however, the application of Pattern Languages is more varied and specific to each use case. As the examples provided by Borchers in [3], each use case has a particular Pattern Language that only applies within the context of each use case. Borchers also expands in the elements that define a pattern by adding context, illustrations, examples and diagrams to the definition of Gamma.

2.1 Interaction Pattern Language

It is clear that the definition and application of a Pattern Language is strongly linked to the context surrounding its application. The basic paradigm of the research described in this paper is that of Cooperative Guidance and Control (CGC) where both the driver and a cooperative automation have influence on the guidance and control of a vehicle, e.g. described in [4–7]. In the case of CGC, Fig. 1 shows an excerpt of the pattern language developed for this. In the diagram, each of the items represents a pattern, with the lines being the links or relation to others. It is important to mention that other possible connections exist but in order to maintain clarity, they have been omitted. An important aspect of a pattern language is that patterns hold a degree of independence; this means that whether a pattern is applied or not, should not affect the implementation of others. This also allows for patterns to be included in other languages.

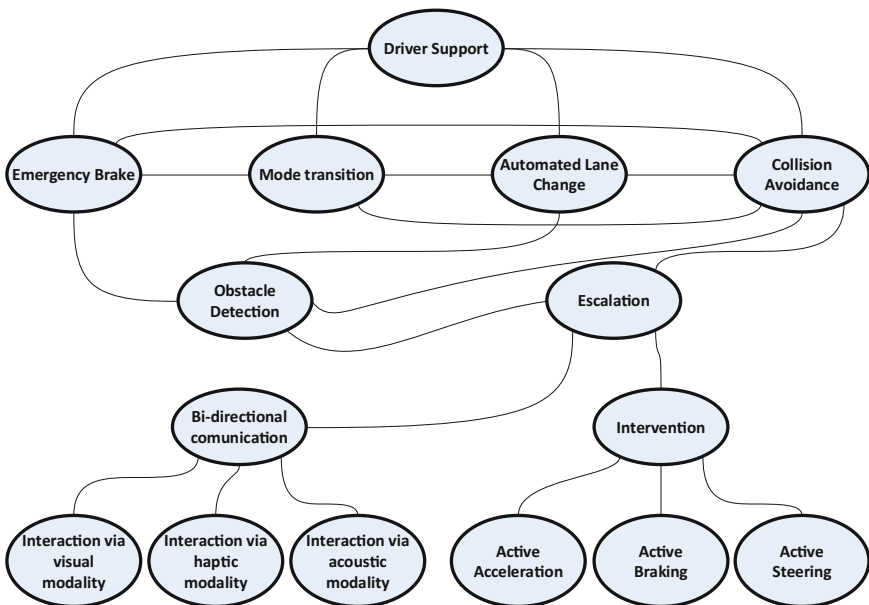


Fig. 1. Diagram of pattern language (CGC)

Each of the patterns in Fig. 1 is described following the same structure. The elements used for this purpose are the following:

Problem. Provides a simplified description of a common problem that the pattern solves.

Solution. Presents a detailed but non-restrictive procedure for addressing the problem. It includes the target domains, image schemas and source domains relevant for solving the problem. In addition, stating the target domains opens the possibility for creativity on how to implement the pattern outside of the proposed image schemas and source domains.

Consequences. Here the possible outcomes of applying the pattern are described. If applicable, affected target domains are also mentioned, e.g. decrease on the sense of urgency.

Implementation Example. These can be programming code examples, technical drawings of a design, or diagrams. Their purpose is to offer an extra explanation and quick guide of how the pattern can be implemented.

For the field Cooperative Guidance and Control it is especially important to point out that the human factors involved in the creation and application of an Interaction Pattern should be included in its definition. This has been done by adding the involved Target Domain or Internal Target States, Image Schemas and Source Domain [8, 9] to the Solution and Consequences elements of the pattern definition.

The Target Domains or Internal Target States are the perception areas that are being changed as a result of the user's interaction with his surroundings. These perception areas, e.g. urgency, importance, authority, etc., can be purposely targeted to generate a desired change by the implementation of an interaction design. Therefore, interaction patterns should consider the Target Domains affected as either part of its solution or as a result of applying it.

The natural learning path since childhood involves getting knowledge by experience; it is during this process that image schemas are built. These image schemas are representations of how our body interacts with the environment and our understanding about it. For example, when used in the design of an interface, image schemas (conventionally written in capital letters) such as BIG-SMALL [8, 10], FAST-SLOW [8], UP-DOWN [11, 12] add greater level of intuitivism. A BIG symbol can be understood as near or more important, a FAST blinking light can mean greater urgency than a SLOW one. We don't explicitly learn the meaning of each of the image schemas but nevertheless we understand them since we have previously encountered them without even realizing it.

As previously mentioned, a Target Domain can be intentionally affected in order to provoke a perception change. Source Domains provide the medium through which new information is captured. They work on the different modalities and have the characteristic of carrying a specific physical value. As such, source domains are closely linked to the implementation of an interaction pattern, e.g. vibration, stiffness, colour, symbols, beeps, etc.

3 Interaction Pattern Example for Shielded Vehicle Application

In this section the application of an example interaction pattern for shielded vehicles will be described. First the respective research questions will be stated followed by the interaction pattern implementation. Afterwards, the study design will be presented. The section concludes with the respective evaluation of the study results.

3.1 Research Questions

The two main research questions of the study were:

Can driving performance and safety in a shielded vehicle with restricted vision in the proximity of the vehicle be improved using interaction patterns?

Can interaction patterns be used to adapt the current driver behaviour towards better performance while reducing the task load?

3.2 Implementation

Name. Collision Avoidance.

Problem. The main problem addressed by the interaction pattern is to prevent collisions with objects in the vehicle's proximity. There are at least two tension poles from the perspective of the ego-system between which a balance needs to be achieved: approaching the obstacle(s) and keeping at distance of (or deviating around) the obstacle. The awareness might be low, perception might be reduced. The problem can be visualized, see Fig. 2.

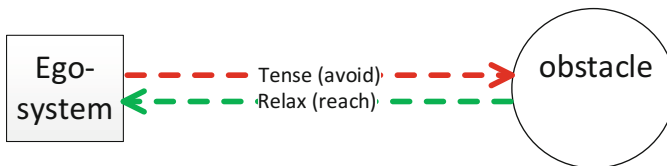


Fig. 2. Abstract problem description for the interaction pattern “collision avoidance”. Upper red arrow represents tensing action due to approaching the obstacle. Lower green arrow represents relaxing action due to receding from the obstacle. The balancing action is to avoid the obstacle and reach a safe distance. (Color figure online)

Solution. A stepwise escalation can lead to a higher awareness that the danger in the current situation is increasing when continuing with the current action or non-action. Also escalation offers the opportunity to react on user action, hence a positive reaction to behaviour adaption towards the non-danger tension pole. Depending on the relative speed and the escalation phase a respective obstacle avoidance manoeuvre should be chosen, e.g. an implemented automated lane change pattern or an implemented

emergency brake pattern. To make both available, a mode transition pattern is necessary to decouple the driver.

The escalation should be triggered depending on the Time to Collision (TTC) in situations with rather high relative velocities, and triggered depending on distance in situations with rather low relative velocities. The result of the pattern implementation is an improved awareness of the target domains *Variation* and *Importance*. *Variation* of the current system state and *Importance* to change the current behaviour.

Consequences. The interaction pattern addresses a reduction in collisions with near objects and accordingly will reduce lateral displacement from the centre of the lane. Due to a reduced task load, the Situation Awareness of drivers will be enhanced. The solution focuses on adaption management, therefore certain internal target states are addressed. Focused domains are *Variation* and *Importance* since an action needs to be made to avoid a dangerous situation that is going to happen in the current course of (non)action. Challenging consequences are the aspect of over trust or overreliance in a non-perfect system. Also connected to non-perfect systems are wrong escalations due to falsely detected objects that may negatively affect acceptance. Finally, an overreaction by the user could be observed when escalation phases are too small.

Implementation Example. The target domain of *Variation* can be addressed using the image schema PATH. Structural elements of PATH are a start, an end and a direction [13, 14]. The PATH schema also includes a series of locations [15] that can be interpreted as escalation steps. Symbolic qualities addressing *Importance* can be implemented with colour codes following the pattern from traffic lights, where red means stop or danger and green means go or safe [16]. These colour codes can be emphasized using the BRIGHT-DARK image schema [12].

The start and end locations or phase limits need to be determined with relevant variables. In a situation with rather high relative velocities, e.g. following a moving car, distance is not sufficient information to determine the need for behaviour adaptation. In situations with high relative velocities the need for reaction is farther away than in situations with low relative velocities. Therefore a combination of relative velocity and distance as time to collision (TTC) seems to be a valid concept to determine escalation boundaries in situations with rather high relative velocities. In more static situations, e.g. parking or slowly driving in a narrow road, the TTC becomes unfeasible due to extremely low relative velocities and the need for very close approaches, so that distance seems to be a useful concept in such situations. Another variable that takes the current driver's behaviour into account is the current steering angle. Therefore depending on the situation the same pattern can be used with different escalation variables and boundaries.

The respective implementations for a collision avoidance pattern in a situation with rather low relative velocities are visualized in Fig. 3 for forward collisions and Fig. 4 for side collisions or road departure.

The PATH image schema is implemented having three distinctive steps or locations: Low urgency, medium urgency and high urgency. The respective variable when a certain urgency level is reached depends on the distance between the ego vehicles bounding box and the obstacle's bounding box as well as the current steering angle. Depending on the ego vehicle's action a *Variation* of *Importance* is defined by the escalation phase.

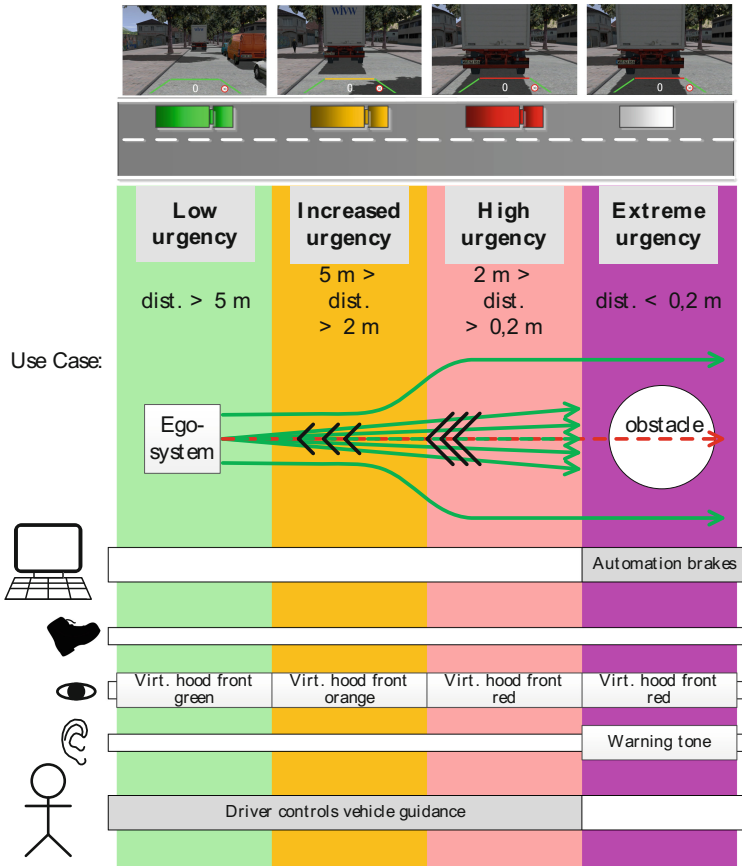


Fig. 3. “Collision Avoidance Pattern” to prevent collisions in forward vicinity. (Color figure online)

In Fig. 3 we see the implementation of the collision avoidance pattern in a forward collision avoidance assistance when approaching a parked truck.

When there is 5 m distance between ego-vehicle and obstacle the forward hood turns from green to orange, except if the driver starts to steer to the right. When forward distance is lower than 2 m and the ego vehicle and obstacle are in the same lane, forward hood turns from orange to red except if the driver starts to steer to the right. If the driver initiates reverse and the distance increases the interaction pattern deescalates, respectively.

A similar collision avoidance pattern can be implemented to be used as a side collision avoidance assistance either for parked vehicles at the side or to prevent road departure (see Fig. 4).

Again not only the current relation of the vehicle towards a side obstacle or roadside end defines the escalation, but also the driver’s action. If the driver already steers to the left when departing to the right, the pattern will deescalate according to the steering model of the vehicle and the respective time of lane departure (TLC).

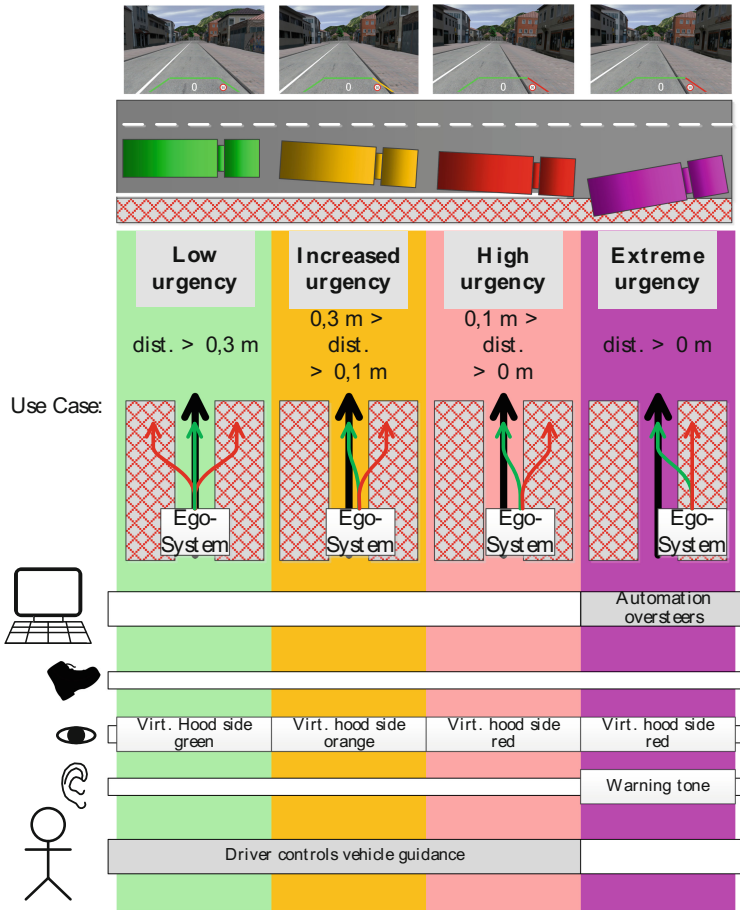


Fig. 4. “Collision Avoidance Pattern” to prevent side collisions or behave like a virtual gravel trap when departing from a road.

3.3 Study Design

The participants drove in a random order with a cabin, a monitor and a monitorsystem with augmentation condition (see Fig. 5).

The design of the experiment was a within subjects design with three repeated measurements. 18 military drivers took part. The mean age of the participants was 32 (SD = 6,3). Every run included a training of 3 min to get used to the setup. After every condition they filled out the NASA-TLX questionnaire. At the end of the experiment the systems were evaluated in a semi-structured interview.

The NASA Task Load Index (NASA-TLX) [17] is an assessment tool that rates perceived task load in order to assess a task. The task load is divided into six subscales. They are rated for each task within a 100-points range with 5-point steps.

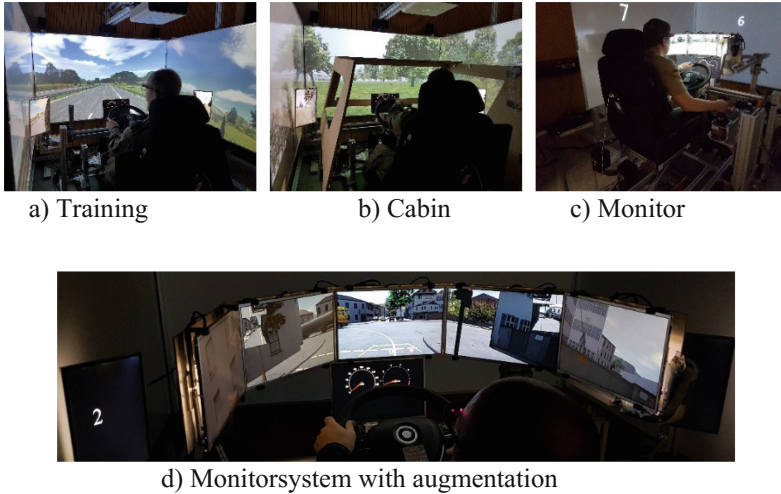


Fig. 5. Test conditions during the experiments

- Mental Demand: How mentally demanding was the task?
- Physical Demand: How physically demanding was the task?
- Temporal Demand: How hurried or rushed was the pace of the task?
- Performance: How successful were you in accomplishing what you were asked to do?
- Effort: How hard did you have to work to accomplish your level of performance?
- Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

Additionally, quantitative data for displacement of the centre of the lane and the amount of collisions with infrastructure, cars and mines was logged during the test runs. Different scenarios were used to find out about the usefulness of the different test conditions in terms of performance and situation awareness (driving data), and task load (NASA-TLX). The respective scenarios were a city scenario with other traffic participants, like pedestrians and other vehicles, and an off-road part. In order to prevent sequence effects, two different maps with different scenario sequences were built and the test conditions permuted.

Respectively, comparable interaction patterns were used in the off-road part, where a very narrow path needed to be followed through a mine field.

Setup. The study was conducted in a generic static driving simulator running the professional driving simulation software SILAB¹. As driving interfaces an active steering wheel and active gas and brake pedals from SENSODRIVE² were used and a Sidestick from Stirling Dynamics³ as gear stick.

¹ WIVW, <https://wivw.de/en/silab>.

² SENSODRIVE, <https://www.sensodrive.de>.

³ Stirling Dynamics, <https://www.stirling-dynamics.com>.

Regarding the visual interfaces, there was a training condition and three test conditions (see Fig. 5).

In the training condition Fig. 5(a), the simulation was visualized via a cave setup representing three large projection screens that were arranged in a 90° angle to the sides and to the front. Additionally two 13" LCD 720p monitors were used as rear-view mirrors. A third 13" LCD 720p monitor was used to visualize speedometer and tachometer.

In the cabin condition Fig. 5(b), a wooden vehicle frame was added to the training condition to introduce ambient occlusion at A and B-pillars.

In the monitor condition Fig. 5(c) and (d), the wooden frame was replaced by a monitor array of five 13" LCD 720p monitors that cover 160° of the driver's horizontal field of view. Rear-view mirrors as well as ramp mirrors were integrated as picture-in-picture (PiP) in the forward left and forward right screens (see Fig. 5(d)).

Additionally, in the monitorsystem with augmentation condition Fig. 5(d), depending on the actions of the driver and the respective situation, interaction patterns escalated or deescalated. As mentioned before, the basic architecture of the generic assistance and automation system is based on the concept of interaction mediation and cooperative guidance and control of highly automated vehicles [6, 7]. In the respective study, only visual assistance is given via the screens. Therefore the final escalation step of the collision avoidance patterns (Figs. 3 and 4), when control was shifted from human to automation, was not considered.

The hypothesis of this study is that the human behaviour adaptation using interaction patterns via a camera monitorsystem with augmentation will improve the situation awareness, the driver's performance and will reduce the overall task load.

3.4 Evaluation

As mentioned before, the concepts were tested in a simulator experiment, in which 18 military drivers took part. The mean age of the participants was 32 (SD = 6,3). All of them have a car and a truck driving licence, eight a motorcycle licence and three a tank licence. 11 of the participants use their vehicle for private purpose daily, 1 participant 3–5 times a week. 50% of the participants had very little simulator experience, 50% little or rather little. Eleven persons assess their driving style as safe/experienced, three as dynamic/sportive/brisk, 4 as cautious. The experience with driving assistance systems, e.g. lane departure warning system, was very low. Only with adaptive cruise control systems 50% of the participants have extensive experience.

In the NASA-TLX (see Fig. 6 and Table 1) the mental demand for driving with the cabin was higher rated than when driving with the monitor and monitorsystem with augmentation condition.

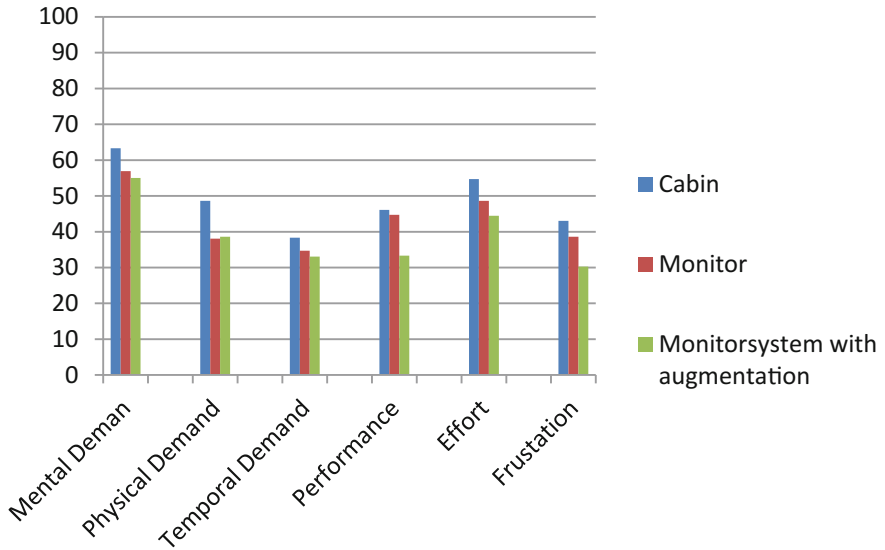


Fig. 6. NASA-TLX diagram

Table 1. NASA TLX scores

	Cabin <i>M (SD)</i>	Monitor <i>M (SD)</i>	Monitorsystem with augmentation <i>M (SD)</i>
Mental demand	63,33 (25,95)	56,94 (24,20)	55 (23,31)
Physical demand	48,61 (21,13)	38,06 (18,16)	38,61 (21,41)
Temporal demand	38,33 (18,15)	34,72 (20,11)	33,06 (15,54)
Performance	46,11* (19,52)	44,72* (19,89)	33,33* (14,65)
Effort	54,72 (23,29)	48,61 (25,19)	44,44 (24,85)
Frustration	43,06 (20,23)	38,61 (23,06)	30,28 (18,90)

* $p < 0,05$

The physical demand was also rated as the highest between the concepts. The monitorsystem with augmentation had the lowest temporal demand, effort and frustration. The participants were satisfied with their performance the most after driving the monitorsystem with augmentation. For the performance there was a statistically significant difference between all conditions. A repeated measures ANOVA showed a difference, $F(2, 34) = 7,055$ $p = .003$, partial $\eta^2 = .293$. A Bonferroni-corrected post-hoc test showed a significant difference between the cabin and the monitor condition ($.009$, 95%-CI $[-22.63, -2.93]$). Also there was a difference between the monitor and the monitorsystem with augmentation condition ($.005$, 95%-CI $[-19.45, -3.3]$).

The lowest lateral displacement (see Fig. 7 and Table 2) was in all parts of the scenarios with the monitorsystem with augmentation. A statistical significance could not be found between the conditions.

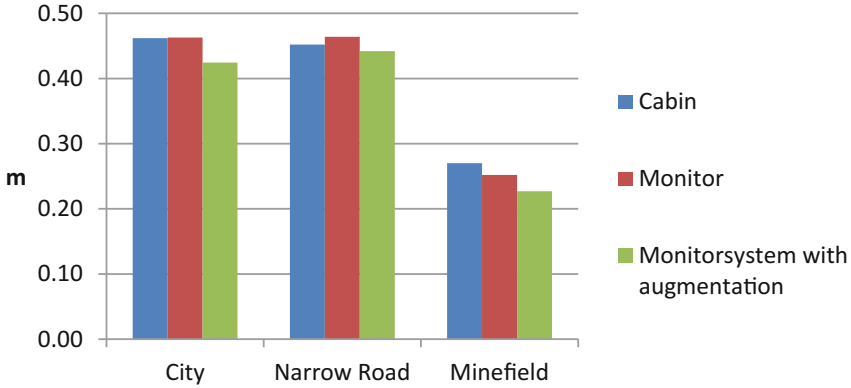


Fig. 7. Displacement from the centre of the lane

Table 2. Displacement from the centre of the lane

	Cabin <i>M (SD)</i>	Monitor <i>M (SD)</i>	Monitorsystem with augmentation <i>M (SD)</i>
City	0,46 (0,14)	0,46 (0,12)	0,42 (0,16)
Narrow road	0,45 (0,07)	0,46 (0,09)	0,44 (0,09)
Minefield	0,27 (0,11)	0,25 (0,09)	0,23 (0,09)

The following diagrams (Figs. 8, 9 and 10) show the different amount of collisions with the respective elements.

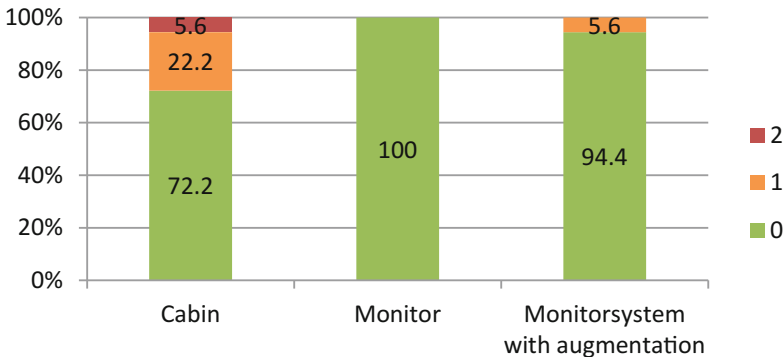


Fig. 8. Collisions with infrastructure

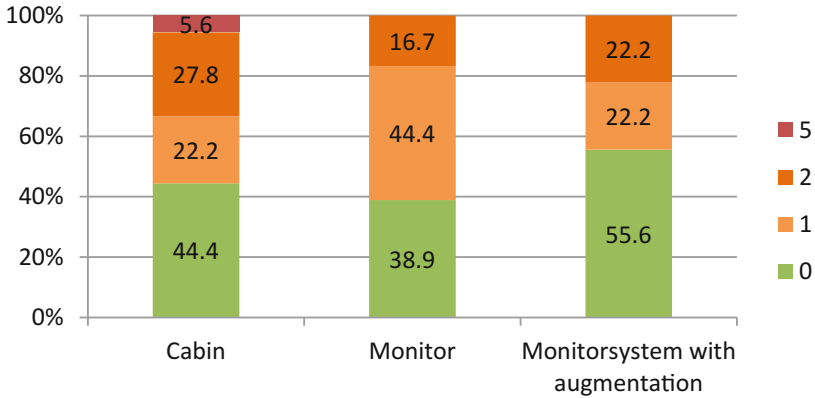


Fig. 9. Collisions with cars

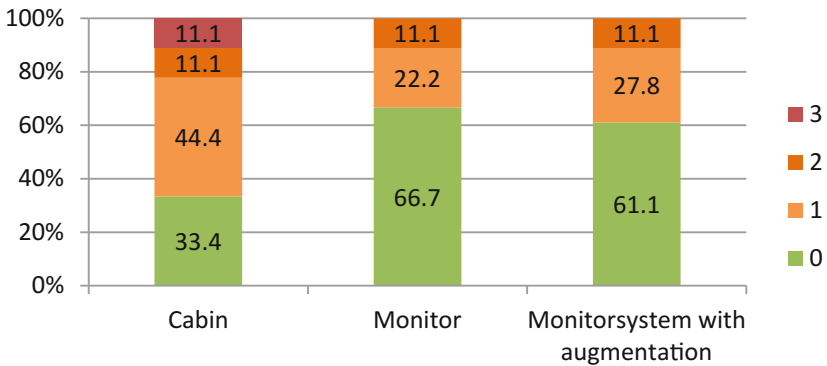


Fig. 10. Collisions with mines

Collisions with infrastructure, cars and mines were mainly caused in the cabin. The probability of a collision is higher, but also the amount of absolute collisions. The infrastructure used in this study was made of houses, walls, traffic lights and traffic signs.

4 Discussion and Outlook

This publication described a systematic approach for behaviour adaptations of humans, based on interaction patterns as a fundamental way to design and describe human machine interaction, and on image schemas as the basic elements of the resulting interaction. This approach was developed in close interplay with the concepts of cooperative guidance and control, where a cooperative automation and a human control a machine together, and of augmented reality, where a natural representation of the world, e.g. in form of a video stream, is enriched with dynamic symbology. The

concept was applied to armoured vehicles, instantiated as interaction patterns “longitudinal and lateral collision avoidance”, implemented in a fix based simulator, and tested with professional operators. The results were a good mix of encouragement and lessons learned, both for the methodical approach of pattern based human machine interaction, and for the application of AR-based cooperative guidance and control.

The use of interaction patterns for collision avoidance, which was tested in the study, showed different aspects of improvement. One of the main objectives to reduce the number of collisions with vehicles in the near proximity could be reached. Also, the displacement was always lower. As a result, the driving performance could be enhanced. Furthermore, the task load could be reduced as the driving with the augmentation caused the lowest temporal demand, effort and frustration. Also, the participants were satisfied with their performance the most.

A possible opportunity for improvement is the integration of live eye tracking to move the drivers focus, e.g. from one screen to another or to point out the critical collision area. Physiological metrics might be useful for patterns in other situations, e.g. combining higher assistance and automation degrees with a mode transition pattern when the task load of drivers is too high. Also, the used visual figure of the hood could be improved in terms of size and form. Additionally, multimodal extension of the patterns with haptic or acoustic feedback could be evaluated.

Regarding the AR-based cooperative guidance and control: We gained an increasing understanding of how this cooperative interplay between an automation and humans can be organized, and patterns are an excellent way to describe this organization. We have encouraging results on a couple of patterns, and especially with the link to image schemas, we increasingly understand why some patterns work differently and better than others. With all optimism, we are far from having optimal patterns, and far from having more than a first glimpse of this vast design space of human machine cooperation and technology based reality augmentation.

Regarding the overall approach of linking patterns with image schemas: For us this is the most promising way to link everything that the community has learned already about specific patterns and specific image schemas, and to make this available in the design process. We have a first understanding how this link can be done, however we are far from having an optimal way to do this linking of patterns and image schemas efficiently. More research, and especially more joint effort is needed to organize and combine the increasing knowledge that is being built up in different spots in the community, and to make this available in the specific design and engineering situation of real products, so that it can improve the increasingly complex human machine systems, not only in the far future, but right here, right now.

References

1. Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., Angel, S.: *A Pattern Language: Towns, Buildings, Construction*. Oxford University Press, New York (1977)
2. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, Boston (1995)

3. Borchers, J.O.: A pattern approach to interaction design. *AI Soc.* **15**, 359–376 (2001)
4. Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., Beller, J.: Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. *Cognit. Technol. Work* **14**, 3–18 (2011)
5. Altendorf, E., Flemisch, F.: Prediction of driving behavior in cooperative guidance and control: a first game-theoretic approach. In: 3. Interdisziplinärer Workshop: Kognitive Systeme. Duisburg, Germany (2014)
6. Baltzer, M.; Altendorf, E.; Meier, S., Flemisch, F.: Mediating the interaction between human and automation during the arbitration processes in cooperative guidance and control of highly automated vehicles: basic concept and first study. In: Stanton, N., Landry, S., Bucchianico, G.D., Vallicelli, A. (eds.) *Advances in Human Aspects of Transportation Part I*, pp. 439–450. AHFE Conference, Krakow (2014)
7. Flemisch, F., Winner, H., Bruder, R., Bengler, K.: Cooperative guidance, control, and automation. In: Winner, H., Hakuli, S., Lotz, F., Singer, C. (eds.) *Handbook of Driver Assistance Systems*, pp. 1–9. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-09840-1_58-1
8. Hurtienne, J.: Image schemas and design for intuitive use – exploring new guidance for user interface design. Dissertational thesis. Technische Universität Berlin (2011)
9. Baltzer, M., Weßel, G., López, D., Flemisch, F.: Interaction patterns for cooperative guidance and control. In: *IEEE International Conference on Systems, Man, and Cybernetics*, Banff, Canada (2017)
10. Tolaas, J.: Notes on the origin of some spatialization metaphors. *Metaphor Symb. Activity* **6**, 203–218 (1991)
11. Lakoff, G., Johnson, M.: *Philosophy in the Flesh: the Embodied Mind & its Challenge to Western Thought*. Basic Books, New York (1999)
12. Baldauf, C.: *Metapher und Kognition: Grundlagen einer neuen Theorie der Alltagsmetapher*. P. Lang, Frankfurt am Main (1997)
13. Johnson, M.: *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press, Chicago (1987)
14. Lakoff, G.: *Women, Fire and Dangerous Things*. University of Chicago Press, Chicago (1987)
15. Lakoff, G.: Some empirical results about the nature of concepts. *Mind Lang.* **4**, 103–129 (1989)
16. Bolz, R.E.: *CRC Handbook of Tables for Applied Engineering Science*. CRC Press, Boca Raton (1973)
17. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Adv. Psychol.* **52**, 139–183 (1988)



Adaptive, Policy-Driven, After Action Review in the Generalized Intelligent Framework for Tutoring

Keith Brawner^{1(✉)}, Alan Carlin^{2(✉)}, Evan Oster^{2(✉)},
Chris Nucci^{2(✉)}, and Diane Kramer^{2(✉)}

¹ Army Research Laboratory, Orlando, USA

keith.w.brawner.civ@mail.mil

² Aptima, Inc., Woburn, USA

{acarlin, eoster, cnucci, dskramer}@aptima.com

1 Introduction and Background

Tutoring on a one-on-one basis from an expert human tutor who is also a subject matter expert represents the ideal arrangement for learning – improving student outcomes on between one and two standard deviations [8]. However, this is not feasible for the vast majority of instruction. One way to attempt attaining similar performance is through the use of Intelligent Tutoring Systems (ITS) – computer systems which can take expert-created content and tutor it with built-in instructional expertise. Systems such as the Generalized Intelligent Framework for Tutoring (GIFT) allow for the creation and configuration of this type of tutoring systems, marrying content from the expert and instruction from a configured system [7].

Human tutors, as opposed to computer tutors, are not statically defined and unchanging – they learn over time. They are able to select key content which focusing on desired learning objectives, and to improve their selections over time after observations of effectiveness [1] – they do what they observe to work. ITS systems should mimic this functionality, by tracking which content sequences teach effectively, and improving content selection and ordering over time. Further, this feedback should be presented in after action review – immediate feedback upon student actions after the student takes them.

The instructional literature indicates that after action review feedback should be focused upon a relatively finite set of immediate learning goals. A brief review of the literature indicates that after-scenario feedback should be:

- Focused - feedback about and at the level of the task step
- Corrective - concerning how to perform specific tasks and steps better (not just feedback regarding accuracy)
- Limited - identify the performance failures with greatest impact and concentrate feedback on them
- Mastery-Focused – feedback should represent deep knowledge or concepts required to master the domain

This short list of feedback items is tied both to Ericsson’s theory of deliberate practice [3] and to Shute’s formative feedback guide [6]. The overall literature presents a picture of beneficial feedback which is short-but-focused and immediately follows an event; at least for the novice learners who comprise the bulk of students. In the design of adaptive intelligent tutoring systems, the above features should be taken into account, but the question becomes how to do so automatically.

One of the primary problems with selecting and improving the content over time is that the data is typically sparse; no two learners are alike, and class sizes are typically less than 100 people. Learners can be clustered into categories, but experience indicates that there are typically less than 5 categories for a desired metric [5], with more metrics resulting in a large number of categories and sparser data. Further, the content can, and usually should, change after each class, creating a “moving target” problem for models. A typical solution of reinforcement learning, where models are learned once and not updated, isn’t an appropriate representation in an arena of both changing content and student populations, which represents a “fluid fitness landscape” in machine learning literature.

An approach to solve this problem involves the creation of artificial students, based on the observed population. These artificial students can have a bell curve distribution of deviations, associated with predictions of how they would experience the content. While the simulated student population does not represent the total population, the approximation is sufficient to create data. This data is then sufficient to create instructional policies. The implementation of the instructional policies is then enough to put into practice. This is especially relevant in relatively sparse selection domains, where the instructional policy is choosing among relatively few pieces of content (e.g. 7 content objects mapping to 3 learning objectives).

This paper describes a “closed loop” system for testing the above design: creating a population of simulated students, approximating the actions that they would take, using those actions to develop a policy of remedial content selection. Further, this paper describes a study which utilizes this technological approach to compare the effect of the developed policies on learners, and reports on the findings.

2 System Design

This system design consists of a few key pieces of experimental apparatus. It is fundamentally built on the architecture of the Generalized Intelligent Framework for Tutoring (GIFT) program – a system of interchangeable software modules for building intelligent tutoring systems. The GIFT modules are the Domain, Learner, Pedagogical, and simulation Gateway. Off-the-shelf versions of these modules were used and configured for the domain of interest. The simulation Gateway was configured to interoperate with the learning environment; the domain module was configured to have a repository of puzzles and per-puzzle feedback. The Pedagogical Module was configured to operate based on an adaptive learning policy, with updates to the Domain

Module to select for the most appropriate feedback. The adaptive learning policy modifications represent the new technical capability presented in this paper, and are described in greater detail below in Sect. 2.2.

2.1 Learning Environment – Physics Playground

Newton’s Playground is a “serious game” – a game designed to teach as well as entertain [9]. During the experience of the game, the learners/players draw different types of physics items, such as a lever, weight, or structural beam. Each of these items is then animated according to the basic laws of physics, and interacts with the rest of the environment – falling, swinging, lifting, etc. as appropriate.

The goal of each level is to get a red ball from one point on the screen to another with a system of drawn objects. An example puzzle may intend to teach the concept of pendulum physics by having a learner draw an anchor, string, and weight in order to have the anchor collide with the ball and launch it to the appropriate screen part.

This work uses a prototype version of Newton’s Playground built for GIFT, where puzzles are instrumented with measures including the completion time for each puzzle. 9 puzzles based on three physics concepts (Impulse, Conservation of Momentum, Conservation of Energy) were constructed using this technology. A sample image of interactions within the environment is shown below in Fig. 1.

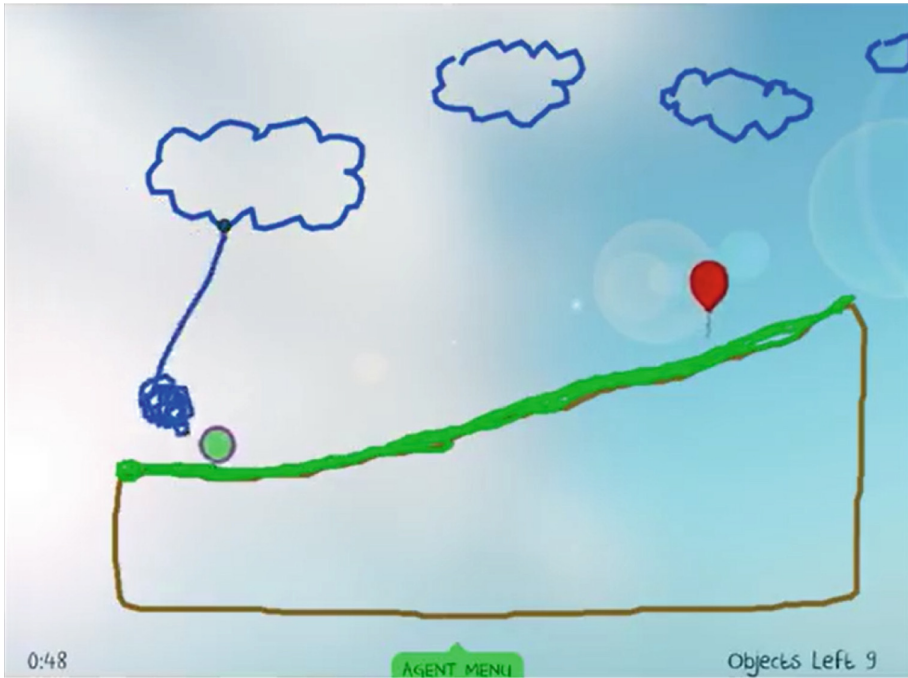


Fig. 1. Physics Playground environment sample interaction; drawing a pendulum to hit the ball uphill to the balloon. (Color figure online)

2.2 Adaptive After Action (A-AAR) Review

The baseline of the A-AAR review is a policy created by the Educational Data Mining (EDM) Tool. The EDM Tool inputs learner performance data and outputs a training model. The input required need only conform to a minimal ontology consisting of learner ID, training item, and at least one measure. At least one of these measures must be specified as a “training goal.” An example of a “measure” is incorrect GIFT report of At/Below Expectation. An example of a system training objective is to guide learners towards “At Expectation” on all associated concepts. Table 1 shows the first few columns of input to the data mining tool from data acquired for this project. The EDM tool augments the data by inferring latent variables. For each row of data, the EDM tool infers a Learner Competency Level (associated with User #123456) and a Scenario Difficulty Level (associated with the Playground and Puzzle associated with the scenario). These values are unknown, but can be inferred by computation, which the EDM Tool fills in using Gibbs sampling [4] (represented in the last three columns). The EDM tool models the process as a Partially Observable Markov Decision Process (POMDP) which serves to maximize the reward (learning) over the observed stats, actions, state transitions, possible observations, and probabilistic models of the

Table 1. Input to EDM tool

Sequence #	User	Playground	Puzzle	Time	Task	Measure 1	Measure 1 Pass/Fail
1	123456	0	0	1/1/2016 13:08	Playground (0): Tutorials	Tutorial	UNKNOWN
2	123456	4	1	1/1/2016 13:10	Playground (3): Conservation of Energy	User draw freeform	PASS
3	123456	3	1	1/1/2016 13:12	Playground (2): Conservation of Momentum	User draw anything	FAIL
4	123456	2	2	1/1/2016 13:14	Playground (1): Impulse	User draw pin	PASS
5	123456	4	2	1/1/2016 13:16	Playground (3): Conservation of Energy	Level completed	FAIL

Table 2. Output of EDM tool (input to pedagogical policy); sample. Below/At/Above Expectation values encoded as 0/1/2. Next is the next node policy.

Node	Puzzle	Lower bound	Upper bound	Next	AAR type
1	Tutorial	0	1	2	
1	Tutorial	1	2	2	
2	Mo 1	0	0	3	Remed
2	Mo 1	1	1	4	Remed
2	Mo 1	2	2	4	Adv
...					

phenomenon. For replication purposes, the exact parameter settings and details of the modeling algorithm can be found within prior work [2]. It is sufficient to indicate that the pedagogical policy is created by an EDM Tool which infers problem difficulties from learning objected measures associated with collected data (Table 2).

2.3 Runtime Integration with GIFT

Previous work sought to integrate Physics Playground with GIFT through traditional GIFT-integration approaches [9]. This combined system is represented in the literature as “Newtonian Talk”, and the development of its system of hints and feedback is beyond the scope of the current work. In this work, the physics environment was used “off the shelf” with built in measures of student assessment and feedback based upon that assessment. The EDM Tool and Policy updates are applied overtop of this existing experience.

The EDM Tool produces Policy updates which serve to update the model of domain content. This model has two functions – first to provide customized feedback from among the pre-authored feedback based on the evidence of which feedback is effective from the policy, and second to provide a recommended puzzle to complete at the next stage. These are presented to the user in the manner of Fig. 3, in the top left and lower right, respectively. In the upper-right of Fig. 3, users classified as “novice”

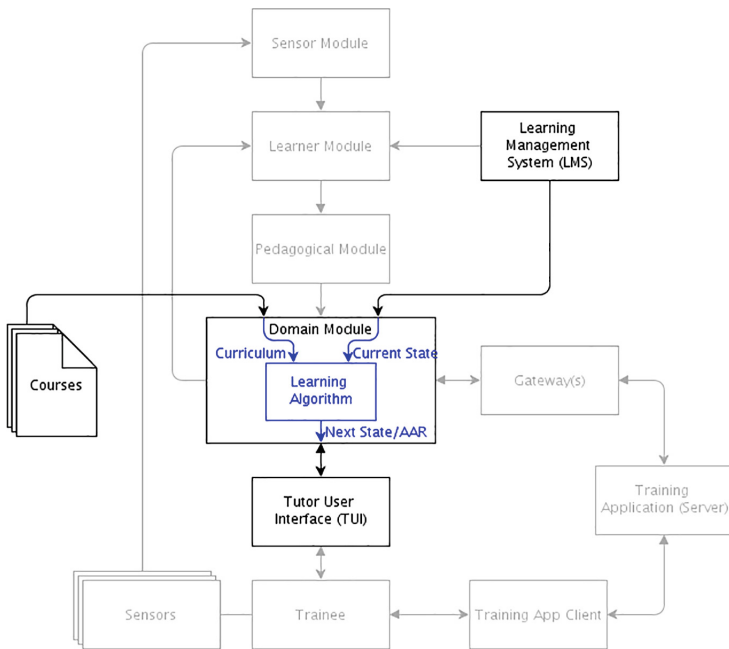


Fig. 2. Learning algorithm implemented into the Domain Module, informed by offline EDM tool processes.

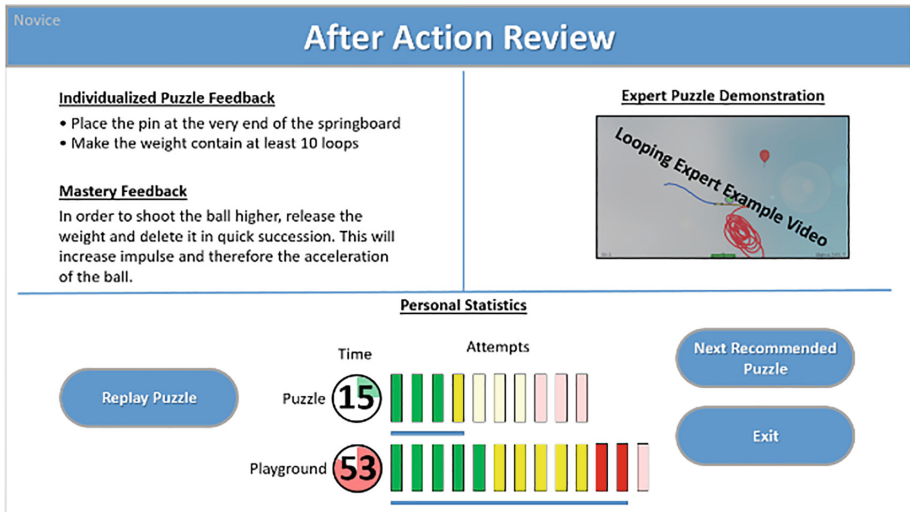


Fig. 3. After-Action Review screen presented to the user.

can play a video of an expert solving the puzzle (presented via an in-window link to youtube), and users classified as “expert” can play a video of their own performance. The feedback screens are integrated into GIFT as shown within Fig. 2, with a customized version of GIFT capable of accepting the policy inputs; this version of GIFT is available upon request of the lead author.

3 Studies and Results

Three different studies were conducted as a part of this project. The goal of the first study was conducted to gather initial data from which to learn adaptive tutoring policies, and to serve as a non-adaptive control condition for future work. In short, without input data, there is nothing for the EDM Tool to generate. The second study used simulated students, with adaptive policy components driving their experiences, and simulated learning outcomes. The goal of the second study was to determine whether the developed policies would have a measurable effect on learning. The third study implemented the developed policies on real individuals, testing the various policies for effectiveness.

3.1 Study 1 - Human Subjects Control, Non-adaptive Policy Baseline

Data was collected on 42 participants, each running through an Introduction to Newton’s Playground lesson followed by 9 puzzles in Newton’s Playground, presented in random order. After the data from each puzzle was recorded, a random policy would select a random next recommended lesson, and the recommendation was displayed to the participant. The 9 puzzles consisted of three puzzles on Energy, three on

Momentum, and three on Impulse. Measures of completion time for each puzzle were recorded. Puzzles took on average about a minute to solve, depending on the puzzle (89, 67, and 82 s respectively for the control conditions of the three puzzles). We found that participants who did not solve the puzzle within that time were likely to take a much longer time to solve the puzzle altogether (floundering behavior). The system timed out after 5 min. To avoid contaminating mean results with the outliers, for data analysis we thresholded performance at 120 s and assigned 120 s completion to any participant who took more than that amount of time. As a sanity check on this thresholding process, we computed median time; for the control conditions it was 99, 64, and 78 s for the three puzzles, and for the experimental conditions it was 63, 25, and 33 s respectively, and thus the difference between conditions was larger for the median than for the mean. Data was stored in the GIFT Learner Management System for later analysis. More complete results within this control condition are presented for comparison in Sect. 3.3.

3.2 Study 2 - Simulated Student Models, Simulated Outcomes

The parameters for the model discussed in previous sections were found by the EDM module run across the initial wave of students in Study 1. There were two types of parameters that were learned from the recorded results. The first type was parameters that were identified through metadata in Newton’s Playground. These included the states (which corresponded to the names of the measures recorded in the LMS), the actions (9 actions, 1 for each puzzle), and the observations (students were recorded for Pass/Fail of the puzzles, with Pass decomposed into AboveExpectation and AtExpectation). The second type included parameters that were approximated through Gibbs sampling of the data [4]. This included the transition probabilities and the observation probabilities for the underlying POMDP model (in other words, the probability that taking each puzzle advances student capability, and the probability of observing a time given student capability, respectively). The observation function was found by finding a best-fit of the item difficulty parameter to the results, based in turn on solving for the student state after each given puzzle.

After the model parameters were determined, a POMDP policy was generated that mapped each state to an action. We simulated 10000 students to validate the policy. The simulation included 10 steps, before the first step, student ability was sampled from a start distribution. Each simulated student iteratively took a puzzle (selected from a Tutorial or the 9 available puzzles), then an observation was received, and then the policy would select the next puzzle. Figure 4 shows an aggregate comparison (averaged over the 10000 students) of this policy to a Non-adaptive policy which selected random puzzles.

3.3 Study 3 - Live Students, Real Outcomes

After the initial data collection, a new study was run using the same pool of 9 N’s Playground puzzles used in the control condition and within the simulated study. This study made the following important changes to the overall protocol: Using of GIFT Cloud, a live system on Amazon Mechanical Turk, to run the subjects; inclusion of

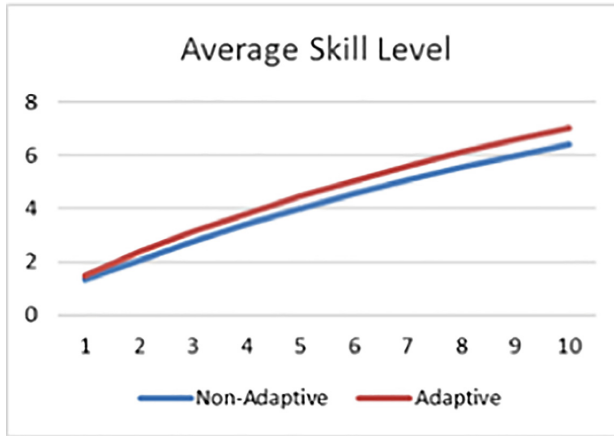


Fig. 4. Simulated student results of differing policies. Greater information on the construction of simulated students is available in the following work [2].

Adaptive AAR screens between puzzles (corresponding to the AAR described in previous sections); lessons were selected using an adaptive policy as opposed to random. Data from the previous data collection was used as a control condition. To provide adequate differentiation between the puzzles selected between the adaptive and control conditions, three puzzles were selected as “test” puzzles (“Impulse 3”, “Momentum 2”, and “Energy 3”, in that order), and the remaining six puzzles comprised the pool of “training” puzzles for adaptation. Of these, four were selected for training, in an adaptive sequence determined by the policy. In summary, each participant in the adaptive condition encountered four puzzles, selected by the adaptive policy, and then was tested on three further puzzles that were withheld from the training pool (Fig. 5).

These individuals are compared against “fair” individuals from the random policies of Study #1. A “fair” individual is one who had encountered at least the same number of puzzles (4, 5, or 6, for the three respective test puzzles, and not including the tutorial) prior to encountering the test puzzle. Correspondingly, the number of fair

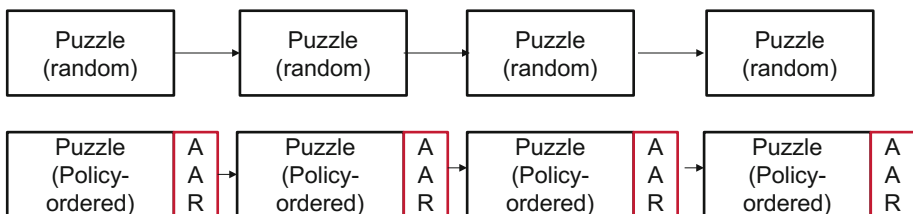


Fig. 5. Control (above) and Experimental (below) conditions. All puzzles have feedback. After 4 puzzles, the experimental condition had a test condition, which is compared against an individual in the control condition who has experienced an equivalent number of puzzles.

Table 3. Completion times of puzzles within the Control (Study 1) and Adaptive (Study 3) conditions, using policies generated with simulated students (Study #2).

Puzzle	Subjects (n = ...)	Control	Adaptive AAR	Absolute delta	Cohen's D	Signif. (p = ...)
Impulse	31 /8	88.23	67.25	20.98	0.60	0.0985
Momentum	19 /8	66.94	35.5	31.44	0.87	0.016*
Energy	17 /8	82.24	45.75	36.49	0.87	0.023*

control individuals shrinks in size when adding criteria. These results are presented within Table 3, with statistical significance marked via asterisk.

4 Discussion

In order for ITSs to be effective, they should closely mirror the functions of the real human tutors that they are modeled after. This includes a continuous analysis of the effectiveness of their actions and the modification of instructional policies accordingly. The above paper presents work performed within the context of GIFT, an intelligent tutoring system with interchangeable parts. A new ‘part’ was designed, developed, prototyped, and tested on live students. This part consisted of an adaptive policy, constructed over top of an existing model of the domain of instruction, which also provided a review/remediation screen.

The adaptive policy was developed on a relatively small sample of real students (<30, on average), but through a relatively large sample of simulated students (1000, exactly). Naturally, a reasonable reader would ask the question of whether a policy developed upon simulated learners would be applicable to live learners, and a follow-up study was run on such learners. This was done with a relatively small sample size, but observed relatively large effects with significance. This comparison somewhat represents the effect of the total intelligent tutoring system, as a policy-driven order of content and a policy-driven AAR page were both introduced. The effect size (large) of intelligent tutoring systems is typically observed to be within the range reported above, and only small samples are required in order to find statistically significant large effects.

The large effect size observed is due to two reasons. The first of these is that the control comparison is against a policy of random actions, but among the same 9 puzzles. However, although the control policy is ‘random’, it does give feedback relative to the learners’ errors. The observed improvements to the system are *in addition to* the basic improvements from error-sensitive feedback. The second reason for the large effect size is that the distribution of the times to solve is bimodal, reflected in relatively high variance values – learner’s either solved or failed to solve the puzzles within 120 s or were considered to have failed and were given feedback/remediation. This bimodal, but clipped, distribution is a fair comparison for real-world application, as the alternative is to allow students unlimited time to flounder and then compare the total floundering times in the control and experimental groups. However, the differences were significant on a unidirectional (we believe the policies should help) test of differences assuming unequal variances.

5 Future Work

The developed models presented within this work are portable, considering that they were made for a system built upon the idea of interchangeable parts. The next steps for this work are to integrate and release the work as open source, as is the practice of the GIFT program. This allows other researchers to benefit from this work at no cost and at no change to their baseline code or models – a powerful advantage. This work is anticipated to be publicly available as part of the GIFT open source package in late 2018, or upon request at time of publication. Another next step is to instructionally test the non-confounded learning items to determine the cause of the learning effect; i.e. is it the intelligent ordering or AAR screen which is causing the learning effect? Finally, the important work in this vein of research is to test the software on a larger sample size, and among different domains. The authors invite collaboration in doing so.

References

1. Carlin, A., Kramer, D., Nucci, C., Oster, E., Freeman, J., Brawner, K.: An adaptive AAR capability for GIFT. In: Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym4), p. 111 (2016)
2. Carlin, A., Nucci, C., Kramer, D., Oster, E., Brawner, K.: Data mining for adaptive instruction. In: Florida Artificial Intelligence Research Society (FLAIRS) Association for Advancement of Artificial Intelligence (AAAI), Melbourne, FL (2018)
3. Ericsson, K.A., Krampe, R.T., Tesch-Römer, C.: The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* **100**(3), 363 (1993)
4. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In: *Readings in Computer Vision* Elsevier, pp. 564–584 (1987)
5. Sachin, R.B., Vijay, M.S.: A survey and future vision of data mining in educational field. In: 2012 Second International Conference on Advanced Computing & Communication Technologies (ACCT), pp. 96–100 (IEEE)
6. Shute, V.J.: Focus on formative feedback. *Rev. Educ. Res.* **78**(1), 153–189 (2008)
7. Sottolare, R.A., Brawner, K.W., Sinatra, A.M., Johnston, J.H.: An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT) Army Research Laboratory (2017). www.gifttutoring.org
8. Vanlehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **46**(4), 197–221 (2011)
9. Zhao, W., Ventura, M., Nye, B.D., Hu, X.: GIFT-powered NewtonianTalk. In: Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym3), Orlando, FL, p. 3 (2015)



Toward Adaptive Training Based on Bio-behavioral Monitoring

Alexis Fortin-Côté¹, Daniel Lafond²(✉), Maëlle Kopf³,
Jean-François Gagnon², and Sébastien Tremblay¹

¹ School of Psychology, Université Laval, Quebec City, QC, Canada

² Thales Canada, Quebec City, QC, Canada

daniel.lafond@ca.thalesgroup.com

³ Thales AVS, Osny, France

Abstract. The present work investigates a cumulative part-task training method that builds up task complexity adaptively based on individual learner states. A research-oriented game entitled “Space Fortress” was used to evaluate two training conditions in a between-group design prior to a third condition involving an adaptive cumulative part task training method. The latter detects when the learner is ready to progress and dynamically adjusts training progression. Here we report the results of the first two conditions. First was the full task condition, where the learner was exposed to the entire task throughout the training session. The second condition followed a cumulative part-task training approach, where sub-tasks were added at fixed progression points. Results showed no statistically significant gain nor loss in terms of learning outcomes between the full task and the non-adaptive cumulative part task condition, adding evidence to previous mixed findings. A trigger rule needed for the adaptive cumulative part task training condition was developed based on short-term patterns of change in performance and mental workload to be used as a dynamic criterion for adaptation. Furthermore, bio-behavioral measures were evaluated as potential proxies for performance and workload with the aim of applying this adaptive method in contexts where performance and workload cannot be directly measured at regular intervals.

Keywords: Adaptive part-task training · Bio-signals monitoring
Serious game

1 Introduction

Learning can sometimes be ineffective due to a training pace that is ill adjusted to individual learners. For complex tasks, reducing cognitive load by splitting the task into sub-components (fractionation) could be beneficial according to cognitive load theory. However, past research [1] has shown fractionation to be ineffective since learning time-sharing skills across sub-tasks is also important. Indeed, cognitive load theory suggests that cognitive load can be divided into three types of load: extraneous, intrinsic and germane, where germane load is associated with the construction and automation of schemas [2]. When learning, particularly complex tasks, the aim of part task training (PTT) is to allow more space for germane load, by reducing the intrinsic

and extraneous load. A way to do it would be to split a complex task into sub-components. There are three different approaches to PTT: segmentation, fractionation and simplification. Segmentation consists in a partition on temporal or spatial dimensions. Fractionation can be used when sub-tasks must be performed simultaneously. Finally, simplification consists in decreasing the difficulty of the whole task by adjusting specific characteristics to it [3].

In addition, for segmentation and fractionation, a specific PTT approach can be addressed with different schedules, i.e. how sub-tasks are combined together. The most common schedules are pure (each sub-task in isolation, then all combined), progressive (two sub-tasks in isolation, then added together), and repetitive/cumulative (one sub-task added to the previous one) [3]. The choice of a schedule has important consequences on the learning process, since it affects the learning of time-sharing skills.

The results regarding the use of PTT are heterogeneous. Studies suggest that PTT was a beneficial method [4, 5], whereas others stated that PTT had no or limited positive effects on learning [1], or was only useful on memory dependent tasks [6]. These results highlight the necessity of considering the approaches separately, as well as taking schedule choices into account when evaluating the efficiency of a method. Moreover, depending on the task on which the experiment is conducted, the results obtained with a specific method will vary.

The present work investigates a cumulative part-task training method that builds up task complexity adaptively based on individual learner states. It is hypothesized that splitting a complex task into sub-tasks and creating a stepwise training session, starting with one sub-task and adaptively adding sub-tasks one-at-a-time, will improve performance in a full-task test. A successful implementation of this method has the potential to improve learning efficiency by reducing training time and/or increasing performance.

In cumulative part-task training (also referred to as repetitive part task training), identifying the optimal trigger for the addition of a sub-task is a key challenge. Indeed, an early trigger might overload the learner, while a late trigger might unnecessarily extend training time. By using two different dimensions, namely workload and performance, we hypothesise that it should be possible to create an integrative rule able to differentiate between an effective learning state and a state where further practice is required. In the present work, six potential rules associated with short-term trends in performance and workload are investigated to select the most promising triggering strategy for adaptive training progression.

Measuring learner progress is another key challenge when quantitative measurements of performance are unavailable. In such cases, physiological measurements might be used as proxy for mental workload and performance, a technique used in previous research during flight simulation [7, 8]. By using task independent metrics focused on bio-behavioural measurements, and by predicting changes in performance and workload, we aim to develop models usable in different training contexts, an approach previously used in different domains such as entertainment technologies [9], emergency management [10, 11] and aerospace [7, 8].

2 Method

2.1 The Space Fortress Game

The present experiment uses a research-oriented video game entitled “Space Fortress” [12] (Fig. 1), a type of serious game [13] which is a task originally developed by cognitive psychologists at the University of Illinois [14]. This task has shown transfer of training to real-world performance in aviation [15]. Wayne et al. [16] studied the acquisition of complex strategies, showing how to capture the explorations, trials, errors, and successes of the learning task in Space Fortress. The aim when playing Space Fortress is to score as many points as possible while controlling a spaceship in an environment where it can be attacked by different opponents (Fortress located at the center of the screen, and mines appearing at random locations). The main task can be split into four distinct sub-tasks that are described in Table 1.



Fig. 1. The Space Fortress game. The ship of the player can be seen in turquoise assaulting the fortress that is always at the center of the screen. A mine can also be seen moving toward the ship of the player. A “\$” symbol can be seen and is part of the ammo management task. The score of the player, ammo, bonus points, and mine indications are displayed at the edge of the screen.

2.2 Experimental Design

The experiment compares two conditions using a between-group design: (1) a full task (FT) condition and (2) a fixed four-step cumulative part-task training (CPT). There are three main types of measurements. First, the game score of each trial is recorded

Table 1. Summary of each sub-task

Sub-task	Description
Navigation	The player has to move his ship to avoid incoming fire from the Space Fortress and being hit by mines
Assault	The player has to aim and fire its weapon in a specific sequence to destroy an immobile Space Fortress located at the center of the screen
Minesweeping	The player has to discriminate friendly from enemy mines depending on their labels, which have to be memorized at the start of the trial
Ammo management	The player has to press either of two buttons when he identifies a predetermined pattern in a continuously changing sequence of symbols on the screen. One of the buttons will recharge ammo; the other gives bonus points

allowing performance comparisons between participants and conditions. Second, ocular, cardiac, and respiratory activity are recorded in order to develop models for inferring the mental state of the participants. Finally, the subjective mental state of participants is assessed using a questionnaire at the end of each trial.

2.3 Participants

Participants were recruited from Université Laval database of volunteers and mailing list. For the FT condition there were 36 participants aged 19 to 46 (M: 25, S: 7) of which 17 were women. For the CPT condition there were 30 participants aged 21 to 50 (M: 26, S: 7) of which 15 were women. The ACPT condition is ongoing with a target of 30 participants.

2.4 Material

Material for all the conditions is identical. It comprised of the game Space Fortress V5 [17], running on a Windows PC, played with a PC controller. A Zephyr BioHarness 3 chest strap, shown on Fig. 2, is used to record the participants' cardiac activity and respiration rate with a raw sampling frequency of 250 Hz and 25 Hz respectively. It also contains an inertial measurement unit reporting acceleration and posture reported at 1 Hz. A Tobii Pro Glasses II eye tracker, shown on Fig. 3, is also used to capture eye movements data. Those include gaze position, pupil size and blink occurrences that are sampled at a frequency of 50 Hz. A Thales developed software is used for synchronising signals and computing advanced features from raw bio-signals such as heart rate variability, eye fixations and spectral density. Logs from the game are used to derive the game metrics such as scores for the different sub-tasks and overall performance on each trial. Finally, participants had to rate after each trial the six statements below selected from previously validated questionnaires [18, 19] on a 5-point Likert scale [20] to record participants' self-reports of engagement and workload.

- I was committed to my goals.
- I have been concerned about achieving my goals.

- It was important for me to perform at this task.
- I have put a great deal of effort into this task.
- I was overwhelmed by this task.
- I was under-stimulated by this task.



Fig. 2. Zephyr™ BioModule sensor and strap (available from BIOPAC.com)



Fig. 3. Tobii Pro Glasses II eye tracker

2.5 Experiment Protocol

For all conditions, the duration of the experiment was approximately 2-h long. Upon arrival, the participants were given a brief overview of the project. The Zephyr BioHarness 3 was installed as well as the Tobii Pro Glasses II. Participants were then asked to read a tutorial to learn the controls as well as the mechanics of the different sub-tasks of the Space Fortress game. They then played four three-minute sessions to familiarize with each sub-task. Participants then completed 24 three-minute trials that differed in the following ways across the three conditions:

1. In the FT condition, participants completed 24 three-minute trials where all the sub-tasks were presented at once, hence they were playing with all four sub-task activated all 24 trials.
2. In the CPT condition, the tasks were cumulatively added at fixed trial numbers. The first 5 trials were comprised of only the navigation sub-task. The trials 6 to 10 were comprised of the navigation, and assault sub-tasks. The trials 11 to 15 were comprised of the navigation, assault and mine sweeping sub-task. The remaining trials were comprised of all four sub-tasks.

The final four trials for each condition are labeled as tests. Indeed, they are identical in all conditions and are used for evaluating learning outcomes.

3 Results

This section presents results from the FT and CPT conditions.

3.1 Learning Rate, Engagement and Workload

First, data was normalized to ensure that conditions could be compared against each other. Normalization is done by participants using the performance score value of the last four trials of the same sub-task group, i.e. the average of the last four trials of comparable score is zero and the standard deviation of the last four trials is unitary for each participant. This is done so that the learning rate can easily be compared between participant and sub-task groups.

A paired t-test on standardized task performance for trials 1–4 vs 20–24 showed a significant effect of the quantity of training sessions on performance in the FT condition (start-end score comparison not possible in the CPT condition), with an average improvement of 2.8 standard deviations, $t(139) = -11.655, p < .001$. Figure 4 shows the normalized score for the FT and CPT condition.

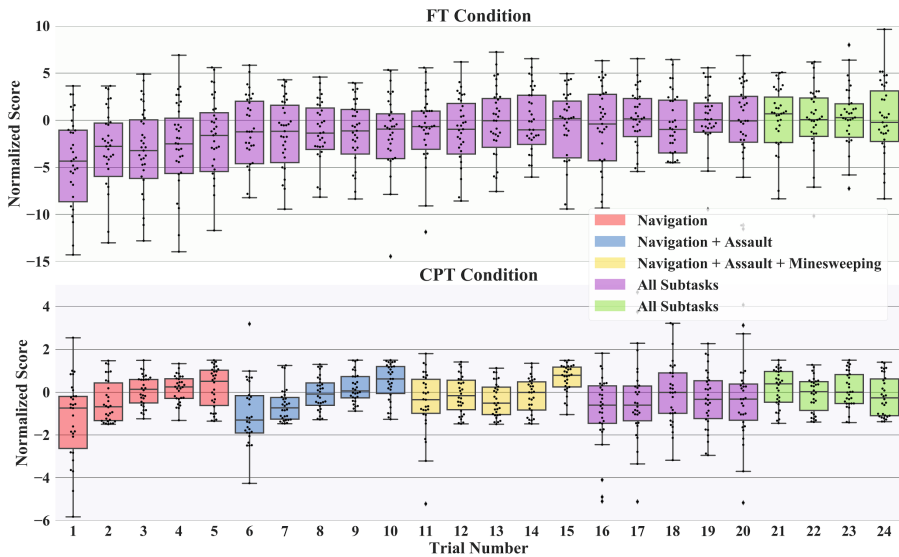


Fig. 4. Performance score for the FT condition and the CPT condition. The different colors represent different sub-task combinations. (Color figure online)

The CPT condition showed performance scores similar to that of the FT condition as shown in Fig. 5. An independent sample t-test on standardized performance showed that the CPT method (M: 0.064, S: 0.945) did not improve learning compared to the FT training method (M: 0.132, S: 0.908), $t(254.66) = -0.595, n.s$. It is important to note that it did not decrease performance either. There is therefore ample room for improving learning efficiency, as is expected to occur in the ACPT condition.

Engagement (on a scale from 0–20) did not significantly differ across the FT (M: 16.73, S: 3.45) and CPT condition (M: 16.69, S: 3.26), $t(1542) = 0.2277, n.s$. Engagement for the final four trials did differ across the FT (M: 16.22, S: 3.98) and CPT condition (M: 17.60, S: 2.98), $t(260) = -3.127, p < 0.01$.

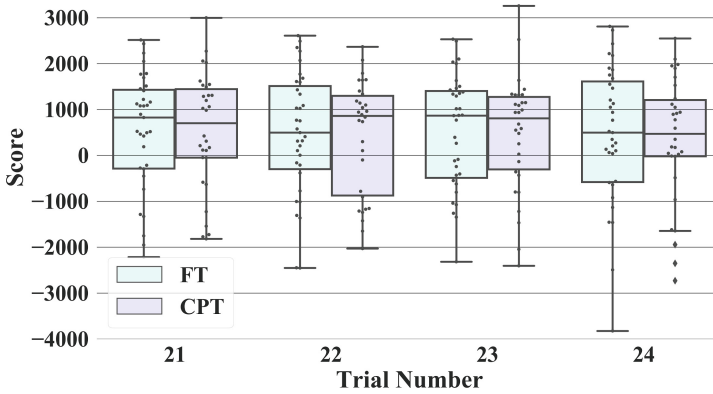


Fig. 5. Final performance scores comparison between the FT condition and the CPT condition.

Workload (on a scale from 1–5) did differ across the FT (M: 2.97, S: 1.43) and CPT condition (M: 3.21, S: 1.31), $t(1542) = -3.419$, $p < 0.001$. Workload for the final four trials did also differ across the FT (M: 3.02, S: 1.32) and CPT condition (M: 3.81, S: 0.99), $t(260) = -5.35$, $p < 0.001$.

3.2 Trigger Rule Selection for the ACPT Condition

By reducing the number of trials in each sub-task group, we hypothesize that it should be possible to reduce overall training time to achieve a comparable performance level to the FT and CPT conditions. To this end, an adaptive training method should be able to detect when the participant has sufficiently learned a sub-task and is ready to move on to the next step cumulatively adding another sub-task. The proposition is that based on performance improvement and workload indication, it is possible to find an optimal trigger rule. Here we investigated a set of six trigger rules that might achieve that. These are, in order of complexity:

1. One trial with a stable or decreasing workload.
2. One trial with an increased performance.
3. Two successive trials with an increased performance.
4. Two successive trials with a stable or decreasing workload.
5. One trial with a stable or decreasing workload and an increase in performance.
6. Two successive trials with a stable or decreasing workload and an increase in performance.

To evaluate objectively the potential effectiveness of these rules, we computed the correlation between the (simulated) number of triggering occurrences for each rule for each participant with their final score on the CPT condition. This condition was chosen because it includes the same sub-tasks and therefore similar difficulty build-up, which will influence workload. The Pearson correlation and corresponding p value is shown in Table 2. As can be observed in the table, only one rule stands out statistically, that is rule number 5. As well as being statistically significant in relation to the final score, this

Table 2. Pearson correlation between triggering occurrences and final score order by the Pearson correlation

Adaptation triggering rule	Pearson r	p value
5. One trial with \searrow workload and \nearrow performance	0.40	0.03
2. One trial with \nearrow performance	0.27	0.14
3. Two successive trials with \nearrow performance	0.22	0.25
1. One trial with \searrow workload	0.15	0.43
6. Two successive trials with \searrow workload and \nearrow performance	0.14	0.44
4. Two successive trials with \searrow workload	0.10	0.60

Note. \searrow means a reduction or no change in the reported workload/performance level since the last trial, while \nearrow means an increase since the last trial.

rule stands out as being a middle ground between triggering too often or not often enough as compared to the other rules which occurs more or less often. Rule 5 was therefore selected as the triggering rule for the ACPT condition.

A simulation of the ACPT condition using the CPT results and the selected trigger rule can be done in order to estimate the training efficiency gain. While remaining a theoretical evaluation (participants still played all the trials instead of progressing early to the next sub-tasks group), the potential efficiency gain can nevertheless be estimated to hypothesize about expected gains in the ACPT condition. Indeed the number of trials that the participants would play can be computed. For the selected rule (5), the distribution of total played trials are presented in Fig. 6. The average number of completed trials is 13 out of a maximum of 20, as in the CPT condition and a minimum of 8 (2 for each sub-task group). There is therefore an average potential saving of seven trials. This represents the best-case scenario for this rule, as this supposes that the participants

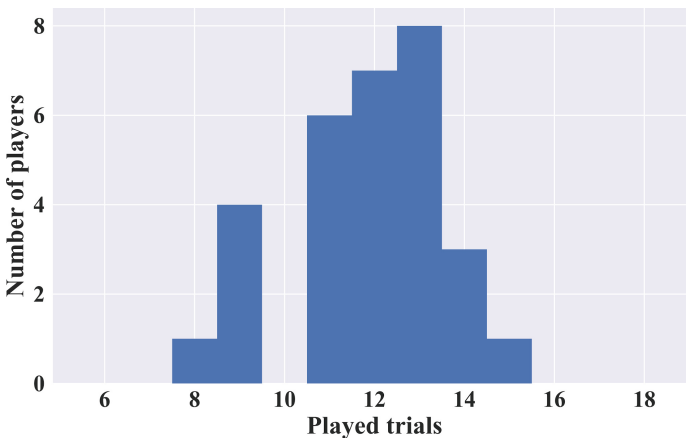


Fig. 6. Distribution of participants by the number of trials played for the simulated ACPT condition according to the rule (5) “One trial with a stable or decreasing workload and an increase in performance”. The maximum possible number of played trials is 20 and the minimum possible number of trials is 8.

are attaining the same level of performance with fewer trials. The real ACPT condition will determine if the participant attained the same level of expertise (final score) as the FT and CPT condition with the lesser number of trials. If they do not exhibit the same level of expertise after their test trials, we will be able to determine at which point they reach the same level as the FT condition as the player will still play the remaining trials for a total of 24 trials. This will allow the computation of the learning efficiency gain of the method by finding the number of trials avoided for the same level of expertise.

3.3 Physiological Measurement as Proxy for Performance and Workload

As a primary sign that at least some of the physiological measurements are indicative of performance, the 7 features with the highest correlation with the score are shown in Table 3. This table also shows in the same way the 7 features with the highest correlation with the reported workload. The following abbreviations are used: Heart Rate (HR), Heart Rate Variability (HRV), Amplitude (Ampl.), Standard Deviation (Std), Acceleration (Acc.). Features are computed on the full-length signal of each trial. Each feature names beginning with the Δ symbol signifies that this value is the difference between the current trial value and the reference resting state value. The HRV Short Window Power Band refers to the spectral power density between 0.05 Hz to 0.15 Hz for windows of 100 s of the Heart Rate signal. The Involuntary Fixation Ratio Long Window is an eye-movement derived feature that is the ratio of time spent under involuntary fixation over the last 60 s. Sagittal, lateral and velocity are features derived from the inertial measurement unit of the Zephyr BioHarness.

Table 3. Pearson correlation between features computed from physiological signals and the trial score ordered by the Pearson correlation

Correlation with score			Correlation with workload		
Feature name	Pearson r	<i>p</i> value	Feature name	Pearson r	<i>p</i> value
Δ HR Max.	0.185	0.000	Δ Sagittal Min. Acc. Ampl.	0.135	0.004
Δ HR Mean	0.160	0.001	Δ Involuntary Fixation Ratio Long Window Max.	0.131	0.005
Δ HR Ampl.	0.157	0.001	Δ Sagittal Peak Acc. Ampl.	0.128	0.006
Δ HR Std	0.154	0.001	Δ Lateral Peak Acc. Ampl.	0.125	0.007
Δ HRV Short Window Power Band Ampl.	0.129	0.006	Δ Pupil Size Average Ampl.	0.120	0.010
Δ Posture Ampl.	0.123	0.008	Δ Mean Velocity Long Window Ampl.	0.115	0.014
Δ HRV Short Window Power Band Std	0.123	0.008	Δ Involuntary Fixation Ratio Long Window Max	0.110	0.019

From this table, it can be observed that Heart Rate features appear to be more correlated with score than body movements and eye movements. Conversely, body and eye movements appear to be more correlated with reported workload intensity. While the correlations are low, they are still significant with p value mostly under 1%.

4 Discussion

In line with previous research [1], the CPT method did not show benefits (nor costs) compared to a baseline FT approach. Self-reported engagement did not change overall but slightly increased for the final four trials between the two FT and CPT condition. Workload increased between the FT and the CPT condition, both overall and for the last four trials. This suggests that the perceived workload was higher for the participant of the CPT condition. This is perhaps because they experienced lower workload in the early phase and therefore affected their reports of the latter, harder, trial workload.

Testing those two methods served as two control conditions to assess the impacts of the ACPT condition. Present results allowed selecting a potentially viable trigger rule for dynamic adaptation, namely a stable or decreasing workload and an increase in performance across two trials. This rule is supported by a positive correlation with learning outcomes (performance in the final test trials), and therefore expected to be successful in detecting the correct moment to trigger the next phase of the training procedure. The ongoing data collection for the ACPT condition will help test the hypothesis that this adaptive procedure may improve training efficiency (either increasing learning outcomes or accelerating the attainment of a same proficiency level). The expected result is that similar score to the FT and CPT condition will be attained in earlier trials, up to an average of seven trials early.

The observed correlations between physiological features and performance score as well as with workload shows promise for training models to detect in real-time performance and workload changes based on bio-behavioural signals. Since different features appear to be correlated with performance and workload, it seems that they may be capturing distinct information about learner state.

Learning retention has not been studied in this experiment. Indeed, each condition might influence learning retention differently over longer periods. Context is also important in adaptive training, while the task presented in this paper is highly controlled with no distractions, training in real-world scenarios might trigger adaptation at inopportune moments if context is not taken into account [21].

Future work includes a second ACPT condition (assuming the first one provides significant benefits), where the trigger rule is based on the output of models built on the bio-behavioural signals instead of the self-reported workload and game score. Indeed not all tasks lend themselves to performance measures and subjective workload ratings throughout a training session. A proxy for those measures based on bio-behavioural signals would therefore make the ACPT method useful for a larger set of training contexts. Inference models have been previously developed for assessing operator functional state [11] a concept that integrates individual human factor dimensions such as workload and stress to assess one's ability to perform current tasks in a nominal fashion [22]. Means for assessing team states have also been proposed [23]. A learner

functional state assessment model could thus be similarly useful in training contexts [24]. As such, the main expected impact of this work is to improve training efficiency in simulators and in the field, in avionics and possibly other related contexts requiring the development of skills and strategies to manage a complex mix of psychomotor, attentional and mnemonic subtasks [25]. Future work will further explore the use of multiple state dimensions, namely workload, performance, engagement and fatigue, to improve the next generation of adaptive training methods.

References

1. Wickens, C.D., Hutchins, S., Carolan, T., Cumming, J.: Effectiveness of part-task training and increasing-difficulty training strategies: a meta-analysis approach. *Hum. Factors* **55**, 461–470 (2013)
2. Sweller, J., van Merriënboer, J.J.G., Paas, F.G.W.C.: Cognitive architecture and instructional design. *Educ. Psychol. Rev.* **10**, 251–296 (1998)
3. Wightman, D.C., Lintern, G.: Part-task training for tracking and manual control. *Hum. Factors J. Hum. Factors Ergon. Soc.* **27**, 267–283 (1985)
4. Johnson, K.B., Syroid, N.D., Drews, F.A., Lazarre Ogden, L., Strayer, D.L., Pace, N.L., Tyler, D.L., White, J.L., Westenskow, D.R.: Part task and variable priority training in first-year anesthesia resident education. *Anesthesiology* **108**, 831–840 (2008)
5. Mané, A.M., Adams, J.A., Donchin, E.: Adaptive and part-whole training in the acquisition of a complex perceptual-motor skill. *Acta Psychol.* **71**, 179–196 (1989)
6. Whaley, C.J., Fisk, A.D.: Effects of part-task training on memory set unitization and retention of memory-dependent skilled search. *Hum. Factors* **35**, 639–652 (1993)
7. Harrivel, A.R., Stephens, C.L., Milletich, R.J., Heinich, C.M., Last, M.C., Napoli, N.J., Abraham, N., Prinzel, L.J., Motter, M.A., Pope, A.T.: Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing. In: *AIAA Information Systems-AIAA Infotech @ Aerospace*. American Institute of Aeronautics and Astronautics, Reston, Virginia, p. 559 (2017)
8. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F.: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* **44**, 58–75 (2014)
9. Mandryk, R.L., Inkpen, K.M., Calvert, T.W.: Using psychophysiological techniques to measure user experience with entertainment technologies. *Behav. Inf. Technol.* **25**, 141–158 (2006)
10. Gagnon, J.-F., Lafond, D., Rivest, M., Couderc, F., Tremblay, S.: Sensor-hub: a real-time data integration and processing nexus for adaptive C2 systems. In: *Proceedings of the Sixth International Conference on Adaptive and Self-Adaptive Systems and Applications* (2014)
11. Parent, M., Gagnon, J.-F., Falk, T.H., Tremblay, S.: Modeling the operator functional state for emergency response management. In: *Proceedings of the 13th International Conference on Information Systems for Crisis Response and Management* (2016)
12. Donchin, E.: Video games as research tools: The Space Fortress game. *Behav. Res. Methods Instrum. Comput.* **27**, 217–223 (1995)
13. Djaouti, D., Alvarez, J., Jessel, J.-P.: Classifying serious games: the G/P/S model. In: Felicia, P. (ed.) *Handbook of Research on Improving Learning and Motivation through Educational Games*, pp. 118–136. IGI Global (2011)
14. Mané, A., Donchin, E.: The space fortress game. *Acta Psychol.* **71**, 17–22 (1989)

15. Boot, W.R., Basak, C., Erickson, K.I., et al.: Transfer of skill engendered by complex task training under conditions of variable priority. *Acta Psychol.* **135**, 349–357 (2010)
16. Wayne, D., Gray, M.: Where should researchers look for strategy discoveries during the acquisition of complex task performance? The case of space fortress. In: Proceedings of the 38th Annual Conference of the Cognitive Science Society, pp. 668–673 (2016)
17. CogWorks CogWorks/SpaceFortress. In: GitHub. <https://github.com/CogWorks/SpaceFortress>. Accessed 22 Nov 2017
18. Leiker, A.M., Bruzi, A.T., Miller, M.W., Nelson, M., Wegman, R., Lohse, K.R.: The effects of autonomous difficulty selection on engagement, motivation, and learning in a motion-controlled video game task. *Hum. Mov. Sci.* **49**, 326–335 (2016)
19. Klein, H.J., Wesson, M.J., Hollenbeck, J.R., Wright, P.M., DeShon, R.P.: The assessment of goal commitment: a measurement model meta-analysis. *Organ. Behav. Hum. Decis. Process.* **85**, 32–55 (2001)
20. Likert, R.: *A Technique for the Measurement of Attitudes*. The Science Press, New York (1932)
21. Fuchs, S., Schwarz, J.: Towards a dynamic selection and configuration of adaptation strategies in augmented cognition. In: Schmorow, Dylan D., Fidopiastis, Cali M. (eds.) AC 2017. LNCS (LNAI), vol. 10285, pp. 101–115. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_7
22. Durkee, K.T., Pappada, S.M., Ortiz, A.E., Feeney, J.J., Galster, S.M.: System decision framework for augmenting human performance using real-time workload classifiers. In: 2015 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision (2015). <https://doi.org/10.1109/cogsima.2015.7107968>
23. Bracken, B.K., Palmon, N., Romero, V., Pfautz, J., Cooke, N.J.: A prototype toolkit for sensing and modeling individual and team state. *Proc. Hum. Fact Ergon. Soc. Ann. Meet.* **58**, 949–953 (2014)
24. Lawrynczyk, A., Chaouachi, M., Lajoie, P.S.: Multimodal assessment of pilots' affective states using psychophysiological sensor signals and facial recognition analysis (2017)
25. Hoke, J., Reuter, J., Romeas, C., Montariol, T., Schnell, M., Faubert, T.: Perceptual-Cognitive & Physiological Assessment of Training Effectiveness (2017)



The Motivational Assessment Tool (MAT) Development and Validation Study

Elizabeth Lameier^{1(✉)}, Lauren Reinerman-Jones^{1(✉)},
Gerald Matthews^{1(✉)}, Elizabeth Biddle^{2(✉)}, and Michael Boyce^{3(✉)}

¹ University of Central Florida, Orlando, FL 32826, USA
{elameier, lreinerman, gmatthews}@ist.ucf.edu

² The Boeing Company, Orlando, FL 32826, USA
elizabeth.m.biddle@boeing.com

³ Army Research Laboratory, West Point, NY, USA
michael.w.boycell.civ@mail.mil

Abstract. The purpose of the present research is to validate a measure of motivation collimated from an individual's motivational, affective, and personality traits. The Motivational Assessment Tool (MAT) is being developed to assess multiple variables for an Intelligent Tutoring System (ITS) to deploy individualized adaptations through various levels of learner profiling. This first study factor analyzed a pool of 303 questions aimed at reducing, refining, and developing scales. Overall, the results of the first factor analysis shows that the MAT is composed of 28 factors. The produced scales are supported by correlations with other factors identified in psychology. The MAT is envisioned to provide inputs into an intelligent tutor's pedagogical strategy to adapt its learning methods to support the learner's motivational type.

Keywords: Motivation · Motivational Assessment Tool
Intelligent Tutoring Systems

1 Introduction

1.1 Motivation and Learning

Motivation, which refers to a student's desires, needs and goals in the educational context, is an important factor in learning outcomes. Motivated students have a drive to succeed that helps them learn [1], while students lacking motivation obtain lower levels of mastery and retention. Motivation has been shown to predict the level of learning goals a student will achieve [2, 3]. Motivation has multiple facets that may shape the learning process in different ways. An important categorization of motivation is intrinsic vs. extrinsic motivation [4–6] and describes the source of an individual's motivation. Intrinsic motivation occurs when the source of motivation is an internal desire to achieve based on the interest and challenge of task performance. With respect to learning, a student with intrinsic motivation has an internal desire to acquire knowledge and explore challenging material. The other category is extrinsic motivation

in which an individual is motivated by external sources such as monetary reward, career advancement or receipt of a certification.

There are many theories and factors pertaining to an individual’s motivation. In general, motivation theories assume that individuals differ in the way that their motivation is affected by environments, including learning environments [7–12]. Reinerman et al. [13] defined a set of motivation variables (see Fig. 1) that contribute to an individual’s motivation. These variables are interrelated such that each individual’s motivation is different. Likewise, each student’s motivation is affected differently to by his or her learning environment. However, identifying an individual’s composition of these variables is difficult because established motivation assessment surveys tend to focus on a limited amount of variables, such as Grit [14] and the 3 × 2 Achievement Scale [15]. Thus, the present research aimed to develop a comprehensive multidimensional assessment, the Motivational Assessment Tool (MAT), through systematic sampling of motivational constructs.

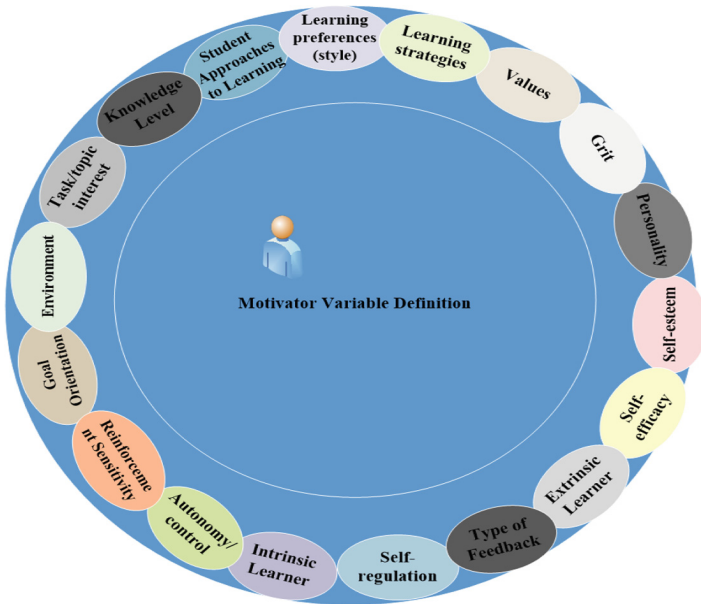


Fig. 1. Motivator variables

In classroom learning environments, instructors can interpret verbal and non-verbal signals from their students to gauge motivation of the class in general. Unless they are able to provide one-on-one tutoring, it is not possible to tailor their instructional strategy to support an individual student’s motivational needs. ITSs have the potential to assess, plan, and implement individualized motivational strategies in real-time to optimize the student’s motivation. The challenge is determining a goodness of fit for each individual that is measured by an increase in effort, attention, goal attainment, learning outcomes, and retention within an intelligent tutor.

The motivational questionnaires, such as the Motivated Strategies for Learning (MSLQ) reviewed under this project were developed with the classroom context in mind. Given the differences between a classroom environment (social, open, competitive) and an ITS environment (private, isolated) some of the questions are not relevant in an ITS environment. Also, new features that an ITS environment brings to the learner are not included in the questionnaires that were built for the classroom, such as the availability of hints, remedial materials and real-time, personalized feedback. Also, traditional motivational questionnaires seek to categorize the learner's motivational traits – relatively fixed characteristics that affect their motivation while learning such as self-regulation and goal-orientation. However, these assessments do not identify the impact of specific items or activities, motivators or reinforcers, that help the learner become or maintain motivated. There are reinforcer inventories built towards school age children and a few for adults, such as the Behavioral Assessment Guide [16] and Dunn- Rankin Reward Preference Inventory [17], but none are used in combination with the general motivation to the complete compilation or for an ITS. Thus, the MAT introduced here will have two sections: a general motivation assessment and the motivator assessment tool for reinforcers.

A blending of both the gaming and classroom needs to occur within the general motivation and the more specific motivators used throughout learning for an IST. This learner profile will begin to shape the relationship needed for the ITS to personalize, adapt, and maintain.

1.2 Motivation Assessment Tool (MAT)

The creation of the MAT began with a literature review in order to identify motivational influences and factors. Then, 31 existing assessments were compiled, overlaps were identified, and constructs were reduced to a basic set. Questions for the scale were arranged by similarities and regrouped until sets of thematically-related questions were sufficiently distinct [13, 18]. This process created the general motivation questionnaire. The motivator inventory was created from existing reinforcer inventories, gaming rewards measures, and definition of new motivators that can be implemented with an ITS. The first MAT had 201 questions for the motivation assessment and 102 questions for the motivator inventory. There were four open-ended questions that could help address any constructs that were missing from the MAT. It was implemented on an intelligent tutoring framework called the Generalized Intelligent Framework for Tutoring (GIFT; 19). Due to the sheer number of items in the MAT, it was broken into 7 sections, 5 for the motivation variables and 2 for the motivator inventory. This iteration of the MAT was then administered as part of a study to support an item analysis for internal consistency and for initial psychometric properties of the scales.

2 MAT Exploratory Analysis

The overarching goal of the study was to develop and refine an initial version of the MAT. This goal was accomplished through three steps. The first was to determine the number of motivational and reinforcer factors present through exploratory factor

analysis. The second was to check psychometric properties of scales that were derived from the factor analysis including internal consistency (Cronbach alpha)'s for the correlations within these scales. The third step was to identify higher order factors from a factor analysis of scale intercorrelations. These steps supported a refined version of the MAT.

2.1 Participants and Procedures

The sample size was 200 participants with ages ranging from 18 through 65 (101 Males and 99 females). There were 3 participants in the 65+ age range, 33 in the 46–64 age range, 71 in the 34–45 age range, and 93 in the 18–33 age range. Participants were recruited by the web platform Amazon Mechanical Turk (AMT). They were provided a link to the GIFT [19], which was used to administer and collect the data. Prior to answering the survey items, the participants read over the consent form and acknowledged their decision to consent. Then, they answered the survey items that included demographic questions and the MAT.

2.2 Measures

The demographic questions asked the participants to report their age range, gender, level of education, income range, GPA. The MAT contained 303 items regarding variables related to student motivation and reinforcers based on the initial motivation taxonomy as described previously. Participants responded to each item on a 7-point Likert scale ranging from “strongly agree” to “strongly disagree”.

3 Results and Discussion

3.1 Initial Item Factor Analyses of the MAT

Two exploratory factor analyses were run to identify underlying constructs for the general motivation and reinforcer sections of the MAT, using the principal factor analysis extraction method. For each analysis, the number of factors was determined from the screen test. The initial factor solution was rotated using the direct oblimin method with Kaiser Normalization, which allows factors to inter-correlate.

14 factors explaining 58.4% of the variance were extracted for general motivation, and 11 factors explaining 64.3% of the variance for motivators. Scales corresponding to each factor were derived from inspection of the factor pattern matrices. Items defining each scale were selected from factor loadings $\geq .5$ where possible, with a minimum loading of .4. A maximum of eight items per scale was chosen.

3.2 General Motivation Scales

Scale Distributions and Alphas. Table 1 lists the 14 MAT General Motivation dimensions identified by factor analysis, together with their working labels. The dimensions fall into three broad thematic groupings consistent with the existing

literature on aspects of motivation. The first group contrasts generally high motivation with vulnerability to loss of focus and interest. The second theme identifies a contrast between positive, approach emotions and social motivations with vulnerability to stress and criticism. The third grouping refers to various aspects of constructive strategy use versus vulnerability to workload.

Table 1. Description of the dimensions for General Motivation

+ Thematic Group: Intrinsic Motivation and Effort —	
<p style="text-align: center;">Learning Driven Interest, hard work, challenge, and persistence directed towards learning</p> <p style="text-align: center;">Goal orientation Motivation to attain performance goals and avoid error</p>	<p style="text-align: center;">Loss of Effort Vulnerability to boredom, lack of focus, and procrastination</p> <p style="text-align: center;">Punishment Energized by threat of punishment to maintain learning</p>
Thematic Group: Social Emotional Factors	
<p style="text-align: center;">Positive Outlook Confidence, optimism, and growth mindset</p> <p style="text-align: center;">Competition Desire to receive a score that is similar or better to peers.</p> <p style="text-align: center;">Social Need for recognition by various individuals</p>	<p style="text-align: center;">Worry Uneasiness towards learning, working with others or making slow progress</p> <p style="text-align: center;">Support Preference Sensitivity to criticism and need for supportive environment</p>
Thematic Group: Task Strategy	
<p style="text-align: center;">Self-regulation Ability to pace learning tasks without direction</p> <p style="text-align: center;">Challenge Need for difficult tasks and complex content to remain engaged in learning</p> <p style="text-align: center;">Breaks Need for pauses throughout learning to focus and energize</p> <p style="text-align: center;">Organize and structure Need to self-structure learning</p>	<p style="text-align: center;">Workload Decrease in performance under high cognitive demands during learning</p>

Scales were constructed to assess the Table 1 dimensions as previously described. Table 2 shows the descriptive statistics of the scales, together with the provisional labels assigned to them, and their alpha coefficients. Cronbach alpha coefficients were generally acceptable: the median alpha was .803 (range: $\alpha = .735-.890$).

Higher Order Scale Factor Analysis. The scales were themselves correlated. Thus, a second-order exploratory factor analysis was conducted, using the same methods as previously. The scree test suggested that three factors should be extracted. Loadings from the pattern matrix $\geq .4$ are listed in Table 3.

Table 2. Descriptive statistics of the MAT General Motivation scales

Type	Scale label	M (SD)	Possible range	Alpha
Intrinsic and effort	Learning driven	61.09 (9.49)	21–77	.890
	Goal orientation	24.69 (5.89)	9–35	.816
	Loss of effort	32.28 (10.80)	10–61	.879
Emotion and ranking	Worry	26.54 (8.63)	7–49	.845
	Competition	18.69 (4.66)	4–28	.803
	Support preference	13.41 (4.91)	4–28	.778
	Positive outlook	22.03 (3.68)	9–28	.781
Task strategy	Self-regulation	65.66 (10.62)	18–91	.864
	Workload	18.98 (6.37)	18–91	.786
	Challenge	27.69 (6.83)	7–47	.735
	Organize and structure	32.35 (3.87)	24–50	.766
Reward orientation	Social	40.46 (10.29)	9–63	.873
	Breaks	12.81 (4.66)	4–28	.762
	Effort based on punishment	12.086 (5.19)	4–28	.811

Table 3. Second order factor analysis of General Motivation scales

Factor	Scales loading on the factor
Factor 1	Workload $-.835$
	Support preference $-.780$
	Loss of effort $-.695$
	Worry $-.688$
	Learning driven $.574$
Positive outlook $.558$	
Factor 2	Competition $.875$
	Social $.640$
	Goal orientation $.606$
	Effort based on punishment $-.461$
Challenge $.410$	
Factor 3	Self-regulation $.615$
	Breaks $.521$

Factor 1 The first factor is a bipolar factor that contrasts two sets of motivational qualities. There were positive loadings for learning driven and positive outlook scales, identifying learners who are intrinsically motivated, confident, and typically hard-working. At the negative end of the scale are learners who are demotivated by workload and prone to lose effort easily. They are also prone to worry and anxiety, and need supportive feedback. Broadly, the factor contrasts optimistic, resilient, and persistent learners with those who are more fragile, avoidant and vulnerable to loss of motivation.

Factor 2 has three major loadings referring to competitive motivations, needs for social recognition, and performance goals. It also has loadings for being motivated by challenge, and a negative loading for punishment. The factor contrasts individuals who see learning as an arena for outperforming others and receiving due recognition with learners who are indifferent to these social motivations and may require some degree of punishment to become motivated.

The third factor was defined by two loadings, both associated with self-regulation. The MAT self-regulation scale refers to motivations to monitor one's learning and take remedial action where necessary. The preference for breaks scale may load on the factor because it refers to the learner's time management skills and capacity to self-regulate focus and alertness.

3.3 Motivator Inventory Factor Analysis

The Motivator Inventory was factor analyzed separately using the same methods described above in the general motivation section. Eleven correlated factors were extracted and rotated, explaining 64.3% of the variance. Items for scale composition were selected on the same principles as before. Working labels and descriptions for the 11 scales are shown in Table 4.

Table 4. Description of the Motivator Inventory Scales

Label	Description
Feedback	Preference for a type and amount of feedback
Recognition	Being acknowledged for your efforts i.e. awards, leaderboards, social media comments, text or emails etc.
Digital	Points, badges, progress bars, a learning companion such as an avatar
Energizer	Use of music, quotes, animated clips, pep talks of avatars to motivate
Logical consequences	A consequence given to the learner due to a lack of effort. i.e. losing points, retaking the course
Low-value	Small prizes such as food, drinks, small gifts, stress ball etc.
High-value	Extra money or promotion
Self-reward	Something you receive i.e. someone cleans your house or subsidized childcare
Activity	Sports that people enjoy i.e. golf
Hobby	Enjoyment of art, theater, concerts, massages
Time	Free time due to efforts and achievement i.e. arriving to work late one day or having a longer lunch

Cronbach alphas were generally acceptable ranging from $\alpha = .670-.920$ (median $\alpha = .876$). Table 2 shows the descriptive statistics of the scales and their alpha coefficients (Table 5).

Table 5. Descriptive statistics of the MAT Motivator scales

Motivator inventory	M (SD)	Possible range	Alpha
Feedback	24.74(6.24)	6–42	.762
Recognition	33.85(11.45)	8–56	.920
Digital	29.55(6.24)	6–42	.902
Energizer	24.03(8.76)	6–42	.896
Punishment	14.85(6.86)	4–28	.914
Low-value	20.27(5.50)	4–28	.858
High-value	31.53(4.76)	11–35	.827
Self-reward	20.60(6.36)	7–35	.670
Activity	33.97(13.88)	10–70	.876
Time after learning	42.82(8.27)	10–56	.840
Hobbies	58.30(18.03)	15–103	.890

Scales were inter-correlated and a second-order exploratory factor analysis was conducted. Three correlated factors were extracted, explaining 68.2% of the variance. Table 6 shows the loadings from the factor pattern matrix that define each factor.

Table 6. Second Order Factor Analysis of the Motivator scales

Factor	Scales loading on the factor
Factor 1	Digital (.774) Energizer (.683) Recognition (.659) Hobbies (.618) Self-Satisfaction (.585) Feedback (.489)
Factor 2	Activity (.728) Exercise (.589)
Factor 3	Time (.780) High value reward (.689)

The Pattern Matrix identified three underlying factors that can be summarized as:
Factor 1: Items related to different types of non-tangible and tangible motivators to include rewards, feedback and interactivity of the learning content

Factor 2: Items that describe types of activities and exercising that can be used as motivators

Factor 3: Items that describe tangible motivators such as money or time to do something the learner enjoys

The first factor explained considerably more of the variance than the remaining two, and correlated with factors two ($\alpha = -.449$) and three ($\alpha = .469$) correlate with the first general factor, which were independent of each other. Thus, the first factor tended to reflect overall sensitivity to reinforcers, but factors 2 and 3 picked up on more

specific influences that cannot be fully separated from the general sensitivity. For example, if a learner is highly motivated by opportunities to play sports (factor 2), this may reflect both general reward sensitivity and a more specific interest in sports.

3.4 Discussion

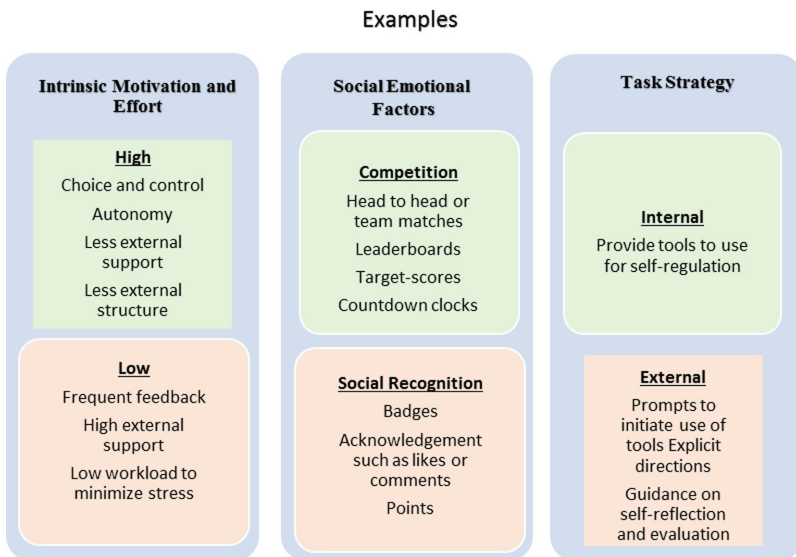
The present research aimed to explore the dimensional structure of learner motivation using an empirical approach based on comprehensive sampling of relevant constructs identified in the literature. It aimed also to distinguish general motivation dimensions associated with the personal characteristics of the learner from dimensions associated with sensitivity to specific reinforcers. Exploratory factor analyses suggested that multiple dimensions of both general motivation and sensitivity to motivators may be distinguished. Motivation dimensions can be thematically grouped in relation to existing constructs, but there appear to be psychometric distinctions that are not well represented in the existing literature. It also proved possible to develop psychometrically acceptable scales for learner assessment on the basis of the factor-analytic findings. Higher-order factor analyses suggest that broader clusters of motivational attributes may also be defined.

Assessment of Motivational Factors in Support of ITS. The first-order factor analysis distinguished a variety of constructs familiar from the research literature e.g., [7, 9, 15]. The three thematic groupings of dimensions identify different types of motivational strengths and weaknesses. Motivation is enhanced by interest in learning and drives to achieve, by sensitivity to social factors, and by self-directed organization of the learning process. Conversely, motivation may be undermined by difficulties in sustaining effort, vulnerability to worry and threats, and poor management of task strategies. All three groupings are relevant to the intelligent tutoring context. For example, three key design issues are (1) how to maintain learner interest and engagement, (2) how to build a sense of being part of a learning community without social threat, and (3) how to enhance the learner's capacities to structure material and utilize effective strategies.

The factor analysis supported development of acceptably reliable scales for general motivation dimensions that can be used to profile the learner's motivational strengths and weaknesses. The second-order factor analysis, based on the inter-scale correlations, revealed a somewhat different factor structure to that suggested by the conceptually-based thematic groupings. Factor 1 pulled in elements of intrinsic motivation and effort, but also emotional concomitants of emotion including a positive outlook, and, at the low end of the scale, excessive worry. It could be seen as contrasting approach and avoidance, seeing these constructs as opposites, rather than independent aspects of motivation as in some theoretical accounts [15]. Factor 2 blends two elements of social motivation – competitive striving against others to excel in performance, and having one's accomplishments be explicitly recognized by others. Factor 3 represents a more narrowly defined version of the task strategy thematic grouping, focusing on self-regulation and taking breaks. Notably, the first-order organization and structure scale did not load substantially on any factor, suggesting that more cognitively-infused elements of task strategy such as note-taking may be distinct from the broad motivational complexes identified in the higher-order analysis.

The present study was directed towards learner characteristics, and assessment of the motivational dispositions that individuals are likely to bring towards a learning environment. Such assessments can then guide personalization of learning. Using the motivation instructional approaches developed in our Phase I research and seen in Table 7 [13], learners that score high on Factor 1 would receive an instructional plan that provides them with control and choices in their learning to feed their intrinsic motivations. Low scorers would follow a plan that minimized excessive workload and stress, and provided frequent supportive feedback. Factor 2 identifies the challenge for online learning environments of limited social interaction. High scorers may require programming of features to support comparison with others and perhaps the ability to communicate with other learners. The role of social motivations also illustrates how discrimination at the finer-grained first-order level can supplement the ‘big picture’ provided by higher-order factor scores. Socially-motivated learners high in competition but not social recognition might require features such as leaderboards, target scores and head-to-head competitions. However, those high on social recognition but not competitiveness might require displays and badges acknowledging their accomplishments without necessarily having to excel against others. Finally, high scorers on Factor 3 might be provided with a variety of aids to self-regulation and refocusing breaks, given that they will be motivated to explore which aids are most helpful to them. Low scorers, lacking motivations to self-regulate, might be given more explicit direction to perform exercises that enhance focus or evaluate progress.

Table 7. Examples of motivational adaptations in an ITS



One of the messages of motivational research is that some learners need more support than others. The general section of the MAT identifies various motivational strengths that will sustain learning even in challenging conditions. A design challenge for online environments, including intelligent tutoring, is how to detect and mitigate lack of motivation or maladaptive motivations such as avoidance. The motivator section of the MAT may be especially valuable in supplementing the general assessment in those cases where motivation is compromised. That is, it may identify the external reinforcers most effective for the individual in those cases where motivation is deficient.

The initial factor analysis of the motivators identified 11 types of reinforcers to which people are more or less sensitive, allowing rewards for participation or achievement to be maximized on an individual basis. The second-order factor analysis of the 11 scales suggested that most of these reinforcers loaded on the first factor. Thus, individuals may differ in general reward sensitivity consistent with psychobiological accounts of variation in the Behavioral Activation System that supports positive reinforcement [22]. However, access to sports and activities, and access to money and leisure time were shown to be somewhat separate classes of reinforcer that may be especially motivating for some individuals. As with the general motivation, assessment of sensitivity to motivators at both first- and second-order levels may support strategies for optimizing motivational support online, especially in those learners that lack intrinsic motivation, competitive drives or effective self-regulation.

Future Research Directions. Identification of missing motivational constructs that are influential for an ITS were constructed to increase the collective depiction of the learner.

The analysis of the subscales produced were discussed and compared to motivational factors to address gaps that may cause a incomplete picture of the learner. Constructs that did not cluster together were added or rewritten. While autonomy is often cited [20] as an attribute of an intrinsic motivated learner, the assessment items in the autonomy section did not cluster together. For our second wave of development and analysis, these questions were reworded to increase the chances of clustering. The Self Determination Theory [21] of an intrinsic learner is founded upon choice and responsibility. The wave 2 analysis will determine if autonomy and level of control will factor within the larger scope of intrinsic motivation. Test anxiety, fear, and fight or flight response with feelings of fear and anxiety were not included and missing constructs for the MAT. The Reinforcement Sensitivity Theory (RST) [22] fit the missing scales and learning questions were constructed around the RST theory scales for the general motivation section.

Motivator Inventory also constructed new scales for addressing additional motivators available for the ITS. A sensor construct was added due to the increasing demand to monitor learners through real-time measures. Sensors can be used as a motivator to increase a drive or for others it may demotivate them due to their sensitivity and anxiety that causes a learner to avoid or shut down. Time was also added in the area of time during learning and negative time. Levels of Interactive Multimedia Instruction (IMI) were added to provide a degree of sensitivity to the type of task a learner prefers from passive tasks, such as PowerPoint to high levels of interaction with

a simulated task. A level of support or frequency and extinction was added to measure an overall level of feedback, points, etc. to motivate the learner. It also provides a general gauge when a motivator may need to be changed because of extinction in order to maintain motivation. Some learners need more frequent positive feedback or points than others. Feedback was divided further to address the different types of feedback and the amount of feedback the learner needs. This study shows that individuals vary in degree of sensitivity to motivational factors. This sensitivity provides multiple pathways for an ITS to adapt and personalize in congruence and extend beyond cognition. The MAT addressed emotional trait characteristics. It seeks to work in congruence with affect, cognition, and traits. Perhaps, there is overlap with trait characteristics that will provide stability for learning motivation and the degree of sensitivity. However, it might be a more fine-grained process divided beyond the categorization of traits. This will be sought after in future validation studies of the MAT.

4 Conclusion

The continuing debate about external motivators and its effect on intrinsic motivation [21] is perhaps, addressed further through the MAT. Specific variables appear to serve as motivators (an external thing that motivates). The correlations for the MAT factor only partially overlapped, which indicates that there may be other motivators outside of external rewards, such as adding challenge to a learning task. Additionally, the results indicate that indeed some individuals are more sensitive to external motivators than others. One general motivator provided is not specific enough to maintain the individual learner properly thus, showing mixed results for external factors when learning. Additionally, perhaps it is more than just one motivator to address an individual needs. A complete system of tailoring externally and internally to achieve a perfect fit, may demonstrate different results. Research generally sticks to one or two controlled variables and not to a more real world messy application of multiple factors in combination that need to be explored due to complexity and furthering our understanding of learning and motivation. Categorizing learners or the fine-grained adaptations is the gateway to determine the true effect of extrinsic variables on an individual. Motivators seem to effect both the driven learner and those that fall on the continuum of extrinsic tendencies as seen in the analysis. Findings have implications for an ITS based on the level or degree of sensitivity or the combination of motivators needed for a learner to perhaps increase learning and retention. The second order factor analysis identified a number of motivation variables found in prior research on motivation and learning, and included in the motivation taxonomy, that can be used to inform the use of a specific instructional strategies designed to support the learner's motivation. In conclusion, an overall comprehensive motivational depiction is comprised of various interrelated variables. These variables affect individuals at different levels of sensitivities that allows for an ITS to deliver adaptive personalized instruction. Adapting to the inter-twinement of variables is an enhancement for optimizing a learner's motivation.

References

1. McCombs, B.L., Whisler, J.S.: The role of affective variables in autonomous learning. *Educ. Psychol.* **24**(3), 277–306 (1989)
2. Ackerman, P.I., Kanfer, R., Goff, M.: Cognitive and noncognitive determinants and consequences of complex skill acquisition. *J. Exp. Psychol. Appl.* **1**(4), 270–304 (1995)
3. Kanfer, R., Ackerman, P.L.: Individual differences in work motivation: further explorations of trait framework. *Appl. Psychol. Int. Rev.* **49**, 470–482 (2000)
4. del Soldato, T., du Boulay, B.: Implementations of motivational tactics in tutoring systems. *J. Artif. Intell. Educ.* **6**(4), 337–378 (1995)
5. Kember, D., Wong, A., Leung, D.: Reconsidering the dimensions of approaches to learning. *Br. J. Educ. Psychol.* **60**, 323–343 (1999)
6. Noels, K., Clement, R., Pelletier, L.: Perception of teachers' communicative style and students' intrinsic and extrinsic motivation. *Mod. Lang. J.* **83**, 23–34 (1999)
7. Ryan, R.M., Deci, E.L.: Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* **25**(1), 54–67 (2000)
8. Alderfer, C.P.: An empirical test of a new theory of human needs. *Organ. Behav. Hum. Perform.* **4**(2), 142–175 (1967)
9. McClelland, D.C.N.: Achievement and entrepreneurship: a longitudinal study. *J. Pers. Soc. Psychol.* **1**(4), 389–392 (1965)
10. Atkinson, J.W.: Motivational determinants of risk-taking behavior. *Psychol. Rev.* **64**(6, Pt.1), 359–372 (1957)
11. Hulin, C.L.: Individual differences and job enrichment—the case against general treatments, Chapter 9. In: Maher, J.R. (ed.) *New Perspectives in Job Enrichment*. Van Nostrand Reinhold, New York (1971)
12. Herzberg, F., Mausner, B., Snyderman, B.: *The Motivation to Work*. Wiley, New York (1959)
13. Reinerman-Jones, L., Lameier, E., Biddle, E., Boyce, M.W.: *Informing the Long-Term Learner Model: Motivating the Adult Learner (Phase 1)*. (Technical report), US Army Research Laboratory, Orlando, FL (2017)
14. Duckworth, A.L., Quinn, P.D.: Development and validation of the short grit scale. *J. Pers. Soc. Psychol.* **92**, 1087–1101 (2009)
15. Elliot, A.J., Murayama, A.J., Pekrun, R.: A 3x2 achievement goal model. *J. Educ. Psychol.* **103**(3), 632–648 (2011)
16. *Behavior Assessment Guide: Reinforcement Inventories for children and adults*, Los Angeles, CA (1993)
17. Catwright, C.A., Catwright, G.P.: Determining the motivational systems of individual children. *Teach. Except. Child.* **2**(3), 143 (1970)
18. Lameier, E., Reinerman-Jones, L., Boyce, M.W., Biddle, E.: Assessing motivation to individualize reinforcement and reinforcers for an intelligent tutor. In: Schmorow, D., Fidopiastis, C. (eds.) *International Conference on Augmented Cognition*, vol 10285, pp. 175–184. Springer, Cham. https://doi.org/10.1007/978-3-319-58625-0_12
19. Sottolare, R.A., Brawner, K.W., Goldberg, B.S., Holden, H.K.: *The generalized intelligent framework for tutoring (GIFT)*, US Army Research Laboratory–Human Research & Engineering Directorate (ARL-HRED), Orlando, FL (2012)
20. Stone, D., Deci, E.L., Ryan, R.M.: Beyond talk: creating autonomous motivation through self-determination theory. *J. Gen. Manag.* **34**, 75–91 (2009)
21. Deci, E.L., Ryan, R.M.: (SDT, n.d.): Questionnaires: Intrinsic motivation inventory (IMI)
22. Mitchell, J.T., Kimbrel, N.A., Hundt, N.E., Cobb, A.R., Nelson-Gray, R.O., Lootens, C.M.: An analysis of reinforcement sensitivity theory and the five-factor model. *Eur. J. Pers.* **21**(7), 869–887 (2007)



A Multi-sensor Approach to Linking Behavior to Job Performance

Alison M. Perez^(✉), Amanda E. Kraft, Raquel Galvan-Garza,
Matthew Pava, Amanda Barkan, William D. Casebeer,
and Matthias D. Ziegler

Advanced Technology Laboratories, Lockheed Martin,
Arlington, VA 22203, USA
Alison.M.Perez@lmco.com

Abstract. Traditionally job performance reviews occur infrequently, only a few times a year at best, and can be largely subjective. Additionally, most quantitative assessment of job performance (e.g., hours at work, number of articles published, etc.) do not give a complete picture, either because they do not account for individual differences or job variability, or they rely only on single measures, subjective reporting, sparse performance measurements or a combination of these factors. Here we report on our initial comparison of objective signals obtained from unobtrusive physiologic and environmental sensors to self-reports of workplace performance and wellbeing. Our results provide evidence that objective metrics of physiological and environmental factors for individuals might be useful in supplementing subjective reports of workplace performance and wellbeing. We posit that a large longitudinal study would provide enough information to automate timely analysis that would allow for tailored performance interventions, workforce retention, and mitigation of negative workplace behaviors.

Keywords: Job performance · Workplace behavior · Physiological monitoring
Wellbeing · Environmental monitoring

1 Introduction

Job performance measurements allow both employers and employees to understand their workplace capabilities over time and can be informative in setting expectations for future workplace behavior and wellbeing. In today's workforce, employees have diverse responsibilities and often face task-saturation [1]. Creating a system that provides multimodal, objective and timely assessments of workplace performance would allow for a comprehensive solution to the challenge of caring for a highly productive but task-saturated workforce. This study is an initial step towards investigating an extensive suite of sensors and metrics in predicting objective and subjective reports of performance and wellbeing. The results of this work will be used for identifying possible connections between signals and self-report assessments as well as being an important building block for planning for and mitigating problems that may occur in larger studies. The lessons learned from this pilot experiment allows us to run a large N

study that can inform changes made to the workplace as a whole as well as in tailoring interventions to individuals that may be impacted by factors of the workplace differently.

Informative metrics of workplace performance and the factors that influence performance and wellbeing can inform interventions and lead to improvements for both employees and employers [2]. Our study compares measures from unobtrusive physiologic and environmental sensors to self-reports of workplace performance and wellbeing (See Table 1). While many studies focus on individual behavioral and physiological measures contributing to individual differences, our sensor suite also includes an often overlooked aspect of job performance by sensing environmental factors that can improve the accuracy of predictions about factors contributing to workplace performance and wellbeing. For instance, ambient CO₂ levels affect cardiac measures, leading to fluctuations in heart rate or HRV that are (1) unrelated to physical exertion or psychological arousal (e.g., Anxiety); and (2) may be independently related to workplace performance outcomes (e.g., reduced higher-order decision-making) [3, 4]. In the first case, adding a CO₂ measurement to the dataset should reduce error in the cardiac-based signals to improve the predictive value of cardiac signals on performance (i.e., convergent validity). In the second case – if CO₂ levels indeed affect decisions influencing workplace performance – adding a CO₂ measurement should also boost the predictive validity (i.e., criterion validity) of the overall performance model.

Table 1. Components of the sensor suite and their relative measurement variables

Sensor	Fitbit	Zephyr	Air Quality Egg	Actiwatch
Signal type	Individual	Individual	Environmental	Individual, Environmental
Signals	Activity Heart rate Sleep quality	ECG Breathing Posture Activity	CO2 conc. Temperature Rel. Humidity	Sleep quantity and quality Natural, artificial, and total light exposure Activity

In this pilot study we identify salient features from our sensor suite that correspond best with ground truth measures of workplace performance, health and wellbeing self-reports through measures of correlation. We will then discuss our plans to use a dual-track approach, combining theory-driven and data-driven models, to link known and unknown connections between sensor signal features to individual job performance variables. These results are intended to support a now ongoing study including 258 participants.

2 Methods

2.1 Participants

Six participants were equipped with an array of sensors, further described below, including: Actiwatch, Fitbit, Zephyr, and an Air Quality Egg over the course of

3 weeks. All participants completed Daily Ground Truth Batteries (DGTB), composed of survey questions on workplace performance, health and wellbeing. All participants were native speakers of English and had at least a high school education. Participants signed the informed consent after reading the experiment summary. The number of daily survey responses per participant ranged from 2 to 21, with an average of 12 responses per participant. The study was approved by and conducted in accordance with the standards of the Western Internal Review Board.

2.2 Sensors

The following suite of sensors was used for this study (For a summary of the sensors, see Table 1).

The Zephyr bioharness is a chest-worn sensor which includes an accelerometer and stretch sensors for breathing rate. It measures three-dimensional postural data and provides ECG (electrocardiogram) and associated heart-related data. While wrist-worn sensors provide some of the same measures, ECG is far more accurate and provides more fidelity than wrist-worn, optical heart rate detection methods [5]. Participants were not required, but had the option to wear the Zephyr bioharness overnight if it did not disrupt their sleep.

The Fitbit Charge 2 is a low-profile wrist-worn activity tracker that can validate many of the same signals as the Zephyr bioharness. It can be worn 24-h to provide measures such as heart rate, sleep quality and other individual measurements of out-of-office activity that may influence wellbeing, health, affect, and burnout. In a recent study comparing wrist-worn trackers to laboratory grade ECG, the Apple Watch, Mio FUSE and Fitbit Charge 2 scored the highest among all sensors tested [5].

The Air Quality Egg is a commercially available environmental sensor that sits on the employee's desk and measures CO₂ concentration in the air, temperature and relative humidity. CO₂ concentration affects HRV and cerebral blood flow increases in response to chronic low levels of CO₂ which in turn can impact executive function and anxiety [3, 6].

The Actiwatch Spectrum Plus is a commercial wrist-worn sensor that is a research-grade sleep quality monitor. Sleep quality (not just time in bed) correlates with physiological health complaints ($r^2 = 0.39\text{--}0.60$) [7]. Also, Sleep and light affect hormones related to overeating behavior [8]. Additionally, there are higher rates of sleep disturbance among those with burnout ($\eta^2 = .396$), and sleep deprivation can affect executive function and anxiety and contributes to deficits in job performance [9–11]. This sensor also detects the light spectrum to allow quantification of an individual's average exposure to different red, green and blue compositions of light. Not only is light exposure linked to depression irrespective of activity levels, but also it is predictive of effect and how seasons and shift-work affect workplace performance [12, 13].

2.3 Experimental Procedure

Participants received the Fitbit Charge 2, Air Quality Egg, the Zephyr bioharness, and Actiwatch Spectrum Plus sensors at the beginning of the study and were trained on how to use all sensors. Participants picked up the Zephyr puck at the beginning of each work

shift at a check-in desk positioned within the facility to maximize convenience and boost throughput during this process. Participants deposited the Zephyr puck prior to leaving work so researchers could upload the data and recharge overnight. Participants who opted in for wearing the device overnight were provided with a second puck when they left work. Participants wore the Actiwatch and Fitbit continuously (24 h a day, for three weeks). Participants also completed daily surveys (DGTB).

2.4 Metrics

The DGTB consisted of various self-report surveys covering topics such as on work-place performance, health and wellbeing. Subsets of the DGTB were administered each day to reduce participant burden. For a summary please see Table 2.

Table 2. Series of questionnaires included in the daily survey batteries

<p><u>Daily Ground Truth Battery</u> In-Role Behavior scale (IRB) Individual Task Proficiency scale (ITP) Organizational Citizenship Behavior (OCB)/Counterproductive Work Behavior Scale (CWB) Big Five Inventory-10 Alcohol quantity Tobacco quantity Physical activity durations Sleep duration Shortened Positive and Negative Affect Schedule Expanded Form (PANAS) Anxiety rating Stress rating Social, Activity and Location Context</p>

2.5 Analyses

Given the limited number of participants, correlations were produced to demonstrate patterns in the data that are of interest. Correlations were derived between daily survey responses and corresponding features from Zephyr, Fitbit, Air Quality Egg, and Actiwatch. Depending on the compliance of each individual wearing each sensor and answering each daily survey, the number of data points differ across each correlation graph. The number of data points is derived the number of surveys taken when each sensor was being used over the course of the 3-week data collection. Features from each of the sensors were averaged over the course of the day of the survey response, except for health survey items, as these items asked about the prior day.

Features from Zephyr were averaged over the course of a day and were then correlated with survey responses ($N = 66$).

Features from Fitbit were averaged over the course of a day and correlated with a subset of the daily survey scores ($N = 61$): Stress, Anxiety, PANAS positive, PANAS negative, and PANAS all (net score), Alcohol, and Exercise. For Alcohol and Exercise, Fitbit data from the prior day is used. All other labels in the correlation plot correspond to Fitbit features.

Features from Air Quality Egg were averaged over 5 ($N = 39$), 15 ($N = 41$), and 30 ($N = 42$) minute windows preceding the submission of daily survey scores for Stress, Anxiety, and PANAS. Features were averaged over 9 a.m.–5 p.m. for Job Survey scores ($N = 21$), and averaged over 9 a.m.–5 p.m. the prior day for Health Survey scores ($N = 30$). Since the Egg only monitors while in the office, weekend data was excluded from the correlations.

Features from Actiwatch were averaged over the day and correlated with the following daily survey scores ($N = 58$): Stress, Anxiety, PANAS positive, PANAS negative, and PANAS all (net score). All other labels in the correlation plot correspond to Actiwatch features.

3 Results

A total of 66 instances of Zephyr features and DGTB responses were correlated across the 6 participants. Some correlations align with the expected direction, concurrent with previous literature, although stronger or weaker given the small number of individuals included in this pilot study [14, 15]. For example, the DGTB item “social_context” options were ordered by decreasing social activity: 1 indicated verbal interactions, 2 indicated written interactions, and 3 indicated no interactions. As expected, there is an inverse correlation with breathing rate, reflecting an increase in breathing rate with speech and reduction with reduced social interaction [14]. However, there are other correlations not expected, such as higher heart rate correlating with an increased number of hours slept. Based on preliminary reviews of this data, this relationship appears to result from a bias of one individual who consistently recorded both higher resting heart rate and the number of hours slept in comparison to the other participants. While raw heart rate was used for these correlations, normalizing heart rates against individual resting rates would likely reduce this type of bias in subsequent analyses. All participants were non-smokers, resulting in a lack of variation in tobacco DGTB responses (See Fig. 1).

A total of 61 instances of Fitbit features and DGTB responses were correlated across the 6 participants. Again, given the limited number of individuals included in the pilot study, our results may not have enough power to show subtle but informative relationships between variables. For example, the Fitbit feature of minutes awake shows a moderate positive correlation with both stress and anxiety, however, given a larger sample we would assume this correlation would strengthen. Other relationships that were expected such as sleep efficiency and exercise showed low to no relation from the correlational analysis (See Fig. 2).

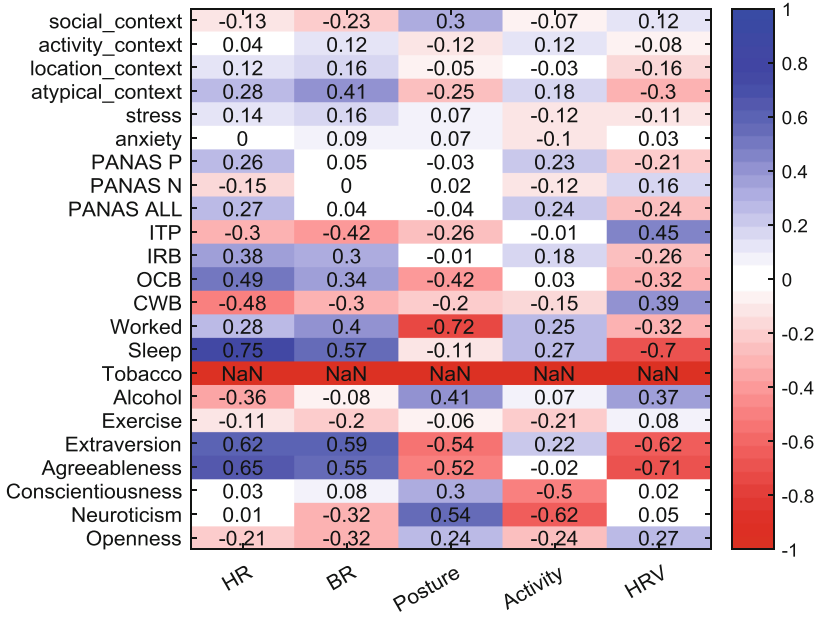


Fig. 1. Correlations between DGTB and Zephyr. NaNs represent that there were no tobacco users in the subject group

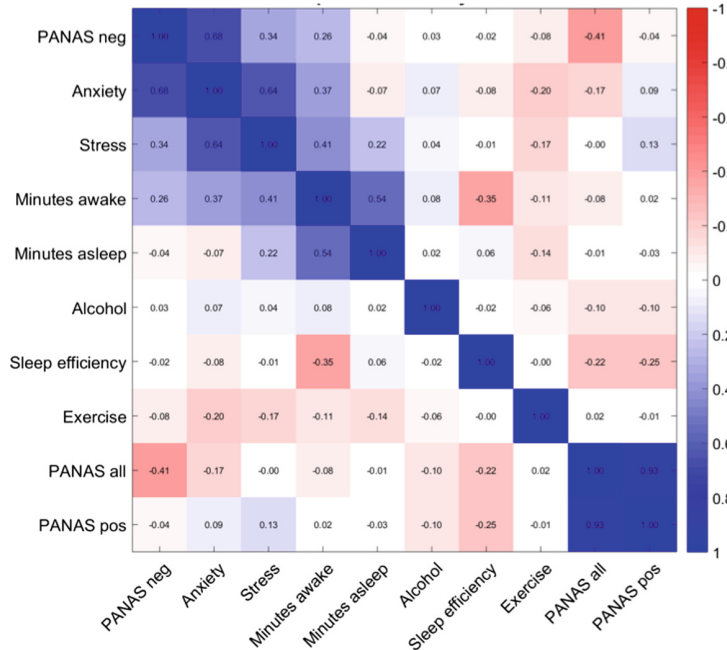


Fig. 2. Correlations between DGTB and Fitbit daily averages

Since the Air Quality Egg only monitors while in the office, weekend data was excluded from the correlations. Labels on the x-axes correspond to daily survey scores, while labels on the y-axes correspond to Egg features. Some correlations were unexpected, such as that between high temperatures and less counterproductive work performance. Some correlations do align with the expected direction, such as the modest correlation between CO₂ composition and higher exercise, and higher CO₂ and lower negative effect (See Fig. 3) [16, 17].

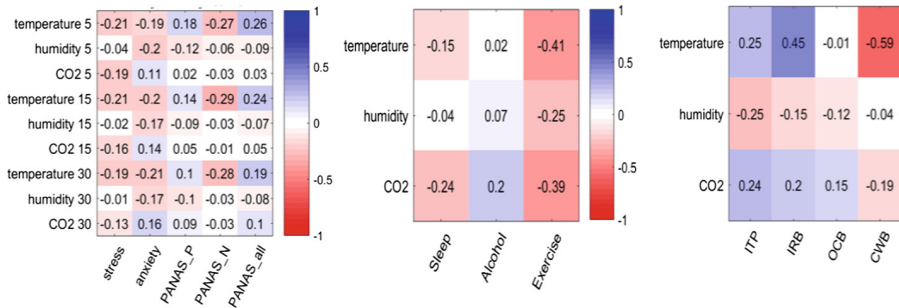


Fig. 3. Correlations between Air Quality Egg and Left: Daily survey items Center: Health survey items and Right: Job survey items

A total of 58 instances of Actiwatch features and DGTB responses were correlated across the 6 participants. High sleep fragmentation, meaning lower sleep quality, was related to higher stress and decreased sleep and immobile time. Sleep fragmentation was also positively correlated with greater light exposure on the same day, including red, green and blue light with blue light having the greatest positive correlation with fragmented sleep as shown in previous research (See Fig. 4) [18, 19].

4 Discussion

In this study, we investigated the individual and environmental factors that may relate to workplace performance, health and wellbeing as measured by ground truth subjective surveys. Correlations presented in this paper are preliminary due to small sample sizes, however this data is useful in identifying possible connections between the signals and self-reports that can be investigated in computational models. We found two important results in this pilot study: First, that even with seemingly motivated participants, the ability to rely on compliance without feedback from the experiment team can result in inconsistent data across sensors and self-report surveys. Second, that the type of data we are recording can be significantly altered by a few individuals that have drastically different physiology or unexpected activities during a day that can throw off any correlations that may otherwise exist. For larger studies it will be important to be proactive in addressing these concerns first by assisting the participants to improve data collection by giving them feedback on a weekly basis on their data

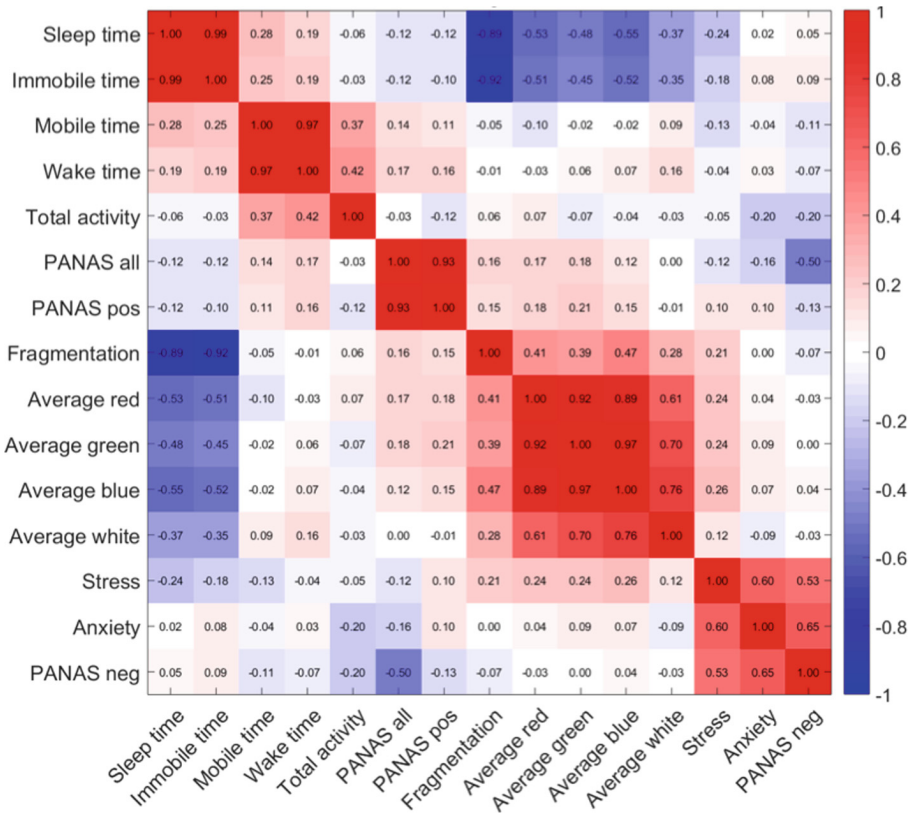


Fig. 4. Correlations between DGTB and Actiwatch daily averages

quality and secondly by using a baseline assessment of each individual throughout the experiment to compare changes over time and personalize the models.

Our future plans are to use a dual-track modeling approach to compare workplace performance variables with the data received from this suite of sensors. The theory-driven modeling component of this approach will allow us to base our predictions on well-established and interpretable relationships between the inputs (signal features) and outputs (job performance and wellbeing), while our data-driven modeling component will investigate novel relationships between signals and job performance metrics not yet represented in current literature. Our theory-driven model will have a fast run time with low computational burden. It will also incorporate new connections between signals and individual variables discovered during the data-driven modeling process.

We have shown in prior work that data-driven models, namely neural networks, can be used to discover new connections by removing constraints between signals and individual variables. As the data-driven approach will include many non-linear combinations of signals that are not easily explainable, we will limit our scope of newly discovered links to those that can be identified by our sensitivity analysis techniques [20].

While the methods we use to derive the data-driven model initially require more computational power, adding newly identified connections to the theory-driven model adds only marginal time and burden to the automated process. We can repeat this technique to add novel sensors and automatically derive new connections between signals and variables to further improve the theory-driven model.

The results of this study and our similar ongoing, larger scale study can lead to a better understanding of employee performance, health and wellbeing in the workforce. Future research should investigate the best methods for using this information to positively impact the workplace. A supervisor, worker, or project team could use the results to assess important workplace conditions with reliable and objective metrics. Future applications of these results could allow individuals and groups to re-engineer their workplace processes so as to enhance performance and productivity while lowering workplace stress.

Acknowledgments. The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17042800004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

1. Jacobs, J.A., Green, K.: Who are the overworked Americans? *Rev. Soc. Econ.* **56**(4), 442–459 (1998)
2. Nbm Alberdi, A., Aztiria, A., Basarab, A.: Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review. *J. Biomed. Inform.* **59**, 49–75 (2016)
3. Pöyhönen, M., Syväoja, S., Hartikainen, J., Ruokonen, E., Takala, J.: The effect of carbon dioxide, respiratory rate and tidal volume on human heart rate variability. *Acta Anaesthesiol. Scand.* **48**(1), 93–101 (2004)
4. Allen, J.G., MacNaughton, P., Satish, U., Santanam, S., Vallarino, J., Spengler, J.D.: Associations of cognitive function scores with carbon dioxide, ventilation, and volatile organic compound exposures in office workers: a controlled exposure study of green and conventional office environments. *Environ. Health Perspect.* **124**(6), 805 (2016). (Online)
5. Wang, R., Blackburn, G., Desai, M., Phelan, D., Gillinov, L., Houghtaling, P., Gillinov, M.: Accuracy of wrist-worn heart rate monitors. *JAMA Cardiol.* **2**(1), 104–106 (2017)
6. Sliwka, U., Krasney, J.A., Simon, S.G., et al.: Effects of sustained low-level elevations of carbon dioxide on cerebral blood flow and autoregulation of the intracerebral arteries in humans. *Aviat. Space Environ. Med.* **69**(3), 299–306 (1998)
7. Pilcher, J.J., Ginter, D.R., Sadowsky, B.: Sleep quality versus sleep quantity: relationships between sleep and measures of health, well-being and sleepiness in college students. *J. Psychosom. Res.* **42**(6), 583–596 (1997)
8. Figueiro, M.G., Plitnick, B., Rea, M.S.: Light modulates leptin and ghrelin in sleep-restricted adults. *Int. J. Endocrinol.* **2012** (2012)

9. Vela-Bueno, A., Moreno-Jiménez, B., Rodríguez-Muñoz, A., Olavarrieta-Bernardino, S., Fernández-Mendoza, J., De la Cruz-Troca, J.J., Bixler, E.O., Vgontzas, A.N.: Insomnia and sleep quality among primary care physicians with low and high burnout levels. *J. Psychosom. Res.* **64**(4), 435–442 (2008)
10. Jackson, M.L., Gunzelmann, G., Whitney, P., Hinson, J.M., Belenky, G., Rabat, A., Van Dongen, H.P.: Deconstructing and reconstructing cognitive performance in sleep deprivation. *Sleep Med. Rev.* **17**(3), 215–225 (2013)
11. McEwen, B.S.: Sleep deprivation as a neurobiologic and physiologic stressor: allostasis and allostatic load. *Metab.-Clin. Exp.* **55**, S20–S23 (2006)
12. Eastman, C.I., Young, M.A., Fogg, L.F., Liu, L., Meaden, P.M.: Bright light treatment of winter depression: a placebo-controlled trial. *Arch. Gen. Psychiatry* **55**(10), 883–889 (1998)
13. Terman, J.S., Terman, M., Schlager, D., Rafferty, B., Rosofsky, M., Link, M.J., Quitkin, F. M.: Efficacy of brief, intense light exposure for treatment of winter depression. *Psychopharmacol. Bull.* (1990)
14. Gupta, J.K., Lin, C.H., Chen, Q.: Characterizing exhaled airflow from breathing and talking. *Indoor Air* **20**(1), 31–39 (2010)
15. Puddey, I.B., Beilin, L.J., Vandongen, R., Rouse, I.L., Rogers, P.: Evidence for a direct effect of alcohol consumption on blood pressure in normotensive men. A randomized controlled trial. *Hypertension* **7**(5), 707–713 (1985)
16. Wasserman, K., Whipp, B.J., Koysl, S.N., Beaver, W.L.: Anaerobic threshold and respiratory gas exchange during exercise. *J. Appl. Physiol.* **35**(2), 236–243 (1973)
17. Woods, S.W., Charney, D.S., Goodman, W.K., Heninger, G.R.: Carbon dioxide—induced anxiety: behavioral, physiologic, and biochemical effects of carbon dioxide in patients with panic disorders and healthy subjects. *Arch. Gen. Psychiatry* **45**(1), 43–52 (1988)
18. Chellappa, S.L., Steiner, R., Oelhafen, P., Lang, D., Götz, T., Krebs, J., Cajochen, C.: Acute exposure to evening blue-enriched light impacts on human sleep. *J. Sleep Res.* **22**(5), 573–580 (2013)
19. Mottram, V., Middleton, B., Williams, P., Arendt, J.: The impact of bright artificial white and ‘blue-enriched’ light on sleep and circadian phase during the polar winter. *J. Sleep Res.* **20**(1pt2), 154–161 (2011)
20. Ziegler, M.D., et al.: Sensing and assessing cognitive workload across multiple tasks. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) *AC 2016. LNCS (LNAI)*, vol. 9743, pp. 440–450. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39955-3_41



Leveraging Cognitive Psychology Principles to Enhance Adaptive Instruction

Anne M. Sinatra^(✉)

US Army Research Laboratory, Orlando, FL, USA
anne.m.sinatra.civ@mail.mil

Abstract. Intelligent Tutoring Systems (ITSs) can be used for computer-based adaptive instruction that can be utilized in many ways including both in the classroom and on a student's own time. ITSs can be particularly useful for remediation and confirming that a student fully understands a topic that is important in an educational course. As many individuals will be using ITSs on their own time, and it is a unique opportunity to customize to an individual, it is helpful to design the material that is being delivered to the student to be as memorable as possible. There are numerous strategies and theories within cognitive psychology that have been heavily researched, and lead to improved memory, and retention when put into place. The current paper discusses how these cognitive psychology strategies can be leveraged and utilized within ITSs in order to lead to improved outcomes. Additionally, there are suggestions on how to incorporate these strategies within ITSs.

Keywords: Intelligent Tutoring Systems · Cognitive psychology
Adaptive instruction

1 Introduction

Intelligent Tutoring Systems (ITSs) provide computer based adaptive instruction to students that can be used as a main instructional tool, an in-class activity, or a way to supplement instruction. An advantage of ITSs is that they are flexible, and customized to the individual student and their instructional needs. It is anticipated that by having a student interact with a highly adaptable computer based learning system which targets the material that they need instruction and remediation on that it will lead to positive learning outcomes and retention of material. While in most ITSs there are many built-in instructional strategies that are intended to enhance retention, there are other less traditional ones that can be utilized in the creation of materials for ITSs, as well as strategies that could be implemented in the future in ITS frameworks.

In cognitive psychology research there have been a number of different activities and principles that have been identified that can enhance the memory that an individual has for information. Among these approaches are the self-reference effect (encouraging the student to link the information to him or herself), heuristics (natural short-cuts and misconceptions that individuals make that lead to quick decisions), mnemonics (memory strategies) and context-dependent memory (improving memory when the context for learning and recall are the same).

This is a U.S. government work and its text is not subject to copyright protection in the United States; however, its text may be subject to foreign copyright protection 2018

D. D. Schmorow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10915, pp. 69–77, 2018.
https://doi.org/10.1007/978-3-319-91470-1_7

The principles and strategies that have been identified in cognitive psychology can be utilized in the creation of learning materials that are provided in an ITS. Many of these cognitive psychology principles can be adapted and applied in a computer based environment to enhance ITSs and ITS frameworks. By incorporating strategies from cognitive psychology into ITSs, they could be used to provide remediation to learners, as well as assist in memory for and retention of learned information.

Examples of leveraging one of these strategies is utilizing the self-reference effect by collecting information about the individual student and storing it in the system as a variable for later inclusion in ITS generated examples and questions. In the current paper different cognitive psychology principles that encourage memory and recall will be discussed, as well as approaches that can be used to harness them in adaptive instruction to enhance human memory and performance. Further, a discussion will be provided identifying different types of memory, and how ITSs as well as their instructional materials can be created to best leverage the way that information is processed.

1.1 Cognitive Psychology

Cognitive Psychology is a subarea of psychology that focuses on how information is processed by humans. One of the defining characteristics that separates cognitive psychology from other areas, is that many of the ideas are abstract and theory driven. For example, we know that the brain retains information, but we do not necessarily know how it does so. There are many theories in cognitive psychology that are used to explain memory such as the modal model of memory, which includes sensory, short term, long term memory [1]. While the different types of memory are generally agreed upon, an area such as attention has more competing theories. For example, memories of attention range from the mechanisms that describe bottlenecks and filters that information need to pass [2], to spotlight theories that use a stage spotlight as a metaphor of what can be focused on [3]. Regardless of the theory that is subscribed to, there are generally “effects” or evidence that is present in the literature that appear to account for improving memory or leading to improved attention. For instance, the “cocktail party effect” is highly tied to attention research. In the cocktail party effect, it has been found that a person’s name can break into attention even when they are engaged in a high workload audio task [4]. This finding can then be applied in new ways such as in emergency alerts [5] or in the cockpit of an airplane. In many ways, cognitive psychology is a discipline that is more abstract, but can be applied in related areas such as human factors psychology or educational psychology.

1.2 Cognitive Psychology and Intelligent Tutoring Systems

As noted, cognitive psychology focuses on the ways that the brain processes information and through research has identified ways that people link information together and process memory. By leveraging the research and “effects” that have been identified in cognitive psychology research it can have a positive impact on learning, particularly in a computer based adaptive system such as an ITS. The approaches can be applied in two ways to assist in learning: (1) as a design principle in the materials and questions

that ITS authors create, and (2) as strategies that are implemented within an ITS framework that will then present material in a way that is consistent with cognitive psychology principles that have been identified as helpful to memory. In the current paper, principles of cognitive psychology that are applicable in an ITS context will be identified, and recommendations will be provided on how they can be utilized to enhance adaptive instruction and ultimately memory for the content that is being taught.

In general, ITSs are made up of four components: a learner module, pedagogical module, domain module, and a tutor-user interface [6]. As an illustrative example, throughout the paper the ITS framework, the Generalized Intelligent Framework for Tutoring (GIFT) will be used to provide context for the recommendations that are provided. GIFT is a domain-independent ITS framework that includes the traditional modules, as well as a sensor module (to allow it to gather data from external sensors) and gateway module (to allow it to communicate to external computer programs) [7]. As GIFT is domain-independent, it provides a structure for the ITS, which is then populated by the course instructor, or ITS author using a set of authoring tools. The ITS author will bring all of his or her own instructional materials and questions to enter into the system for use during tutoring. GIFT is a research project, and is constantly being updated. Therefore, any of the suggestions provided in this paper could be applied in future iterations of GIFT. For instance, the pedagogical module within GIFT provides remediation material and strategies based on a literature review [8]. Strategies that are identified from cognitive psychology could be used to update the types of strategies that are used by the pedagogical module in GIFT.

2 Identified Cognitive Psychology Principles

While there are numerous cognitive psychology principles that exist and can be utilized to improve memory, the current paper will discuss a sample of them and provide examples on how they can be particularly tied into computer-based adaptive instruction. The discussion will begin with an explanation of common theories of memory.

2.1 Theories of Memory

Modal Model of Memory. According to the Modal Model of Memory [1] there are three types of memory: sensory memory, short-term memory, and long-term memory. Sensory memory exists in both the visual and auditory form for an extremely short amount of time. If the sensory memory is paid attention to, then it can transfer to short-term memory, which is where memory is temporarily stored prior to it being moved to long-term memory, where it is retained by the individual. In general, it is believed that rehearsing, or working with the information, will assist in moving it from short-term to long term memory. Some of the evidence for this is found in the serial position effect [9]. If an individual is provided with a list of information, they have a tendency to recall the information based on the order in which it was provided to them. For instance, the information about the beginning of the list is well remembered

because it had an opportunity to be repeated over and over again by the individual (the primacy effect; [9]). Additionally, the information at the end of the list is also remembered by the individual since it is still in short term memory (the recency effect; [9]). Combined, the primacy and recency effect make up the serial position effect.

The serial position effect is an example of a strategy that can be utilized in both the design of ITS material, and within an ITS framework. If an author identifies that the material to be learned is highly repetitive or list based, then the ITS system could intentionally be designed to tutor the middle items or steps in a sequence to ensure that they are recalled to the same degree as the others. For a more generalized approach to harnessing the way that memory works, it would be helpful for the ITS author to keep in mind that working with and processing information in short-term memory can ultimately move it to long-term memory. Therefore, it would be helpful to author materials that include reminders of information that was previously learned, and provide checks on learning that require the learner to use information that they have recently learned.

Additionally, one way to approach facilitating the use of appropriate strategies in an ITS would be to provide the author with an authoring tool interface that asks questions about the characteristics of the material to be taught. For instance, if the author is asked if the material is “list-based” or if “order matters”, then perhaps the ITS could select this as an appropriate approach for them to use and auto-populate the information that is entered by the author in the system.

Working Memory. Baddeley’s model of working memory [10] builds on the Modal Model of Memory. According to Baddeley’s model, as opposed to having a static short-term memory store, humans actually have an active short-term memory that is actively used for thinking and processing of information. It has been shown that the capacity of working memory is approximately 7 plus or minus 2 items [11]. This research has been applied in the form of the length of phone numbers. Baddeley describes working memory as being made up of three components: the visuospatial sketchpad (which deals with visual information), the central executive (which determines what needs to be paid attention to/allocates resources), and the phonological loop (which deals with auditory information) [12]. Research that supports this model has shown that individuals have difficulty processing two types of visual information at the same time, but can process visual and auditory information simultaneously [12]. This suggests that they are separate processes, and that when designing a task it can be helpful to separate visual and language based tasks to have maximum attention and retention. An example of the application of working memory in the real world is, that in driving research that has found that talking on the cellphone while driving impacts driving performance [13]. One explanation for this is that the visuo-spatial sketchpad is being double taxed by the visual driving task, and the visual mental imagery task that is related to the conversation that is occurring.

Per the application of this research, when designing content in an ITS, it may be better to design instructional materials such that any interactions do not double tax the same working memory resource. For instance, if the individual is being tutored in a simulated driving environment it would be preferred to provide tutoring feedback in auditory form as opposed to visual form, and to focus on simple topics as opposed to

language that elicits visual images in the learner's mind. If both the task and feedback are in visual form it may pull the attention of the individual from the task, and it may overload them. In this situation the auditory channel is not yet being used, so it may have better results, especially if the materials provided are straightforward. In an ITS authoring tool it could ask the author about the characteristics of the task, and provide recommendations based on the type of task. For instance, if it is an auditory task, it may recommend visual feedback, and vice versa.

2.2 The Self-reference Effect

The self-reference effect has shown that if information is tied to the self it is easier to recall than if it is not [14]. By integrating strategies and functionalities into an ITS such that it can reference the individual by name it may be able to leverage this effect for better learning [15]. Sinatra and colleagues designed a tutor for learning how to solve logic grid puzzles, and varied the types of names that appeared within the content of the material during learning [16]. Participants either received the interactive tutorial with their own name and the names of friends (to encourage self-reference), names of popular culture characters or generic names included in the text of the tutorial and puzzle that was being used for learning. The idea behind this was that by utilizing one's own name in the material they would potentially learn the material more deeply and be able to transfer their skills more successfully. Interestingly, individual differences played a role with those who were high in need for cognition performing as expected in the conditions when the manipulation occurred during learning (although it was not significantly significant), with best performance with self-reference, and worst performance with generic names. However, those who were low in need for cognition actually performed better when popular culture names were used, and worst when their own names were included. This is an important finding, as it appears that individual differences such as need for cognition can impact how effective or non-effective a strategy is when it is applied.

Providing a way for authors to include the names of individuals within materials could have a positive impact, especially for students who are high in need for cognition. One approach to being able to harness the self-reference effect, would be to use an approach within the ITS that stores a variable, such as the learner's name. When the learner enters their name into a survey in response to a question, the system can then store the variable and reference it later on in material, questions, and instructions. This can result in learners feeling that the material is more personalized, as well as leveraging any potential self-reference effect that exists. It may be helpful to have options with the authoring tools that provide question types that the author can use to create variables that can later be reused in questions and throughout the tutoring instance. Further, in the case of GIFT the self-reference effect could be implemented as a strategy that is recommended when learners score high on a need for cognition scale, but not those that score lower on it.

2.3 Heuristics

Heuristics are rules of thumb that individuals use to help them make quick decisions. However, the use of heuristics do not always result in the correct answer. There are many different common heuristics that individuals use, but for the purpose of this paper two will be discussed in regard to their relationship to ITSs. Heuristics such as the framing effect and the sunk costs effect can be applied to ITSs in different ways. According to the framing effect, the way that a question is phrased will influence the response that an individual provides. For instance, if a question is formed in terms of gain, individuals are more likely to agree that it is the right choice than if it is phrased or “framed” as a loss [17]. The sunk costs effect is also relevant in terms of an ITS in a different way. According to the sunk costs effect, if an individual has already spent time or money on something, they are more likely to continue on with it even if it no longer makes sense [18].

With the framing effect in mind, an ITS author or instructor should be mindful of the way that he or she phrases questions that are being used during the tutoring session or for assessment that will lead to remediation. It is important to differentiate between if the learner does not understand the question or if the question is phrased in a way that will naturally lead the learner to provide a specific answer. This is particularly relevant when tutoring and assessing a topic such as logic, which may ask the learner to choose between two or more answers. If the phrasing of the answers is as a gain or loss it may impact their response. The framing effect may result in individuals consistently providing the wrong answers, and going through unneeded remediation in the tutor. Therefore, it is important to avoid common reasoning errors that individuals may make during the authoring of tutor assessments.

The sunk cost effect can be applied in different ways in ITSs. If an ITS is structured such that learners have an opportunity to select their own topics and provide the order that they will be engaging with the material in there is the potential for them to engage with the sunk costs effect. For instance, if a student is repeatedly misunderstanding and failing questions about a specific concept, the tutor may continually push them back in through remediation and not allow for them to move on until they have been successful in completing it. Since the learner has spent a great deal of time with the system he or she may feel that it would be advantageous to continue pursuing mastery of the topic, but this may be at their detriment because they are unable to move on to additional material that they may be able to master. Therefore, it is important to be mindful of the way that an ITS is structured such that it will not necessarily keep pushing learners through the same material if they are not achieving it, and it may provide reminders to learners to continue on to other topics if they are by choice continually repeating a topic.

2.4 Mnemonics

Mnemonics are strategies that can be used by individuals to help them remember information. Mnemonics can be as simple as creating an acrostic from the first letter of the words that need to be remembered, or as elaborate as remembering items by associating them with specific locations and taking a mental walk through a familiar

environment to recall the list (the method of loci). Research has suggested that using the method of loci in a virtual environment has similar results as using it in a conventional familiar real environment [19]. Similar to the method of loci is pairing two images together and imagining them when it is time to recall them. For instance, if the individual needs to remember an iguana and a book, it may be helpful to imagine the iguana sitting on a book.

If the task that is being tutored is heavily memory based, then the ITS author may want to pair different images together to assist the learner in remembering the items. This can be achieved in an ITS through authoring materials that show the necessary images together. Further, it can be tied to procedural knowledge and route knowledge by tying the step that the individual needs to engage in with another interesting image which can then serve as a retrieval cue to assist with their memory. For instance, if the learner is being taught about how to navigate a path in the environment he or she can be prompted to imagine other items or images in the locations where they need to make turns. Using the iguana example, it may be helpful to show a drinking fountain with an iguana on it, followed by a staircase with a turtle on it. Then when the individual navigates the actual environment within a simulation he or she can recall where the important turns are based on recalling the interesting item that they had previously associated with it.

2.5 Context Dependent Memory

The idea of context dependent memory is rooted within cognitive psychology research. It has been found that when information is recalled in the same setting it is learned, then there are better results [20]. For instance, based on this idea, if an exam was given in the same room that the lecture occurred in, students would likely perform better on recall than if the exam was given in a different room.

This can be applied in ITSs by providing customization for the pages that are used for assessment. For instance, if the author provides the information on a page with a blue background, then can also choose to have their recall/exam on a page with a blue background, which would provide the same context that the material was learned in for the student.

3 Recommendations for Leveraging Cognitive Psychology Principles in Intelligent Tutoring Systems

In the current paper, principles and effects identified within the cognitive psychology literature have been highlighted. Additionally, the connections that these principles have to ITSs has been discussed. There are two main ways that these can be incorporated into ITSs, (1) through the authoring of material and structuring of associated authored questions by the course author, and (2) through implementing features and authoring tools within ITSs that allow for these strategies to be used. Certain principles

lend themselves better to item 1, while others lend themselves more to item 2. In addition, it might be helpful to incorporate features and authoring tools into ITS frameworks such as GIFT to allow for these different principles to be harnessed by ITS authors who may not be familiar with them.

3.1 Initial Suggestions for Authoring Tools to Support the Use of Cognitive Psychology Principles in ITSs

As mentioned earlier in the paper, there are a number of different cognitive psychology principles that may have a positive impact on learning, but are not familiar to an ITS author or course instructor. One way to assist them in being able to use these strategies would be to create an optional ITS authoring tool or component of the existing authoring tool that an author can use to enhance the tutor that they are creating by using cognitive psychology principles.

The specific authoring tool could ask questions about the material that is to be learned, and based on those responses it could provide recommendations on approaches that can be used to author the material. For instance, if the author indicates that they are teaching vocabulary and definitions, then it can recommend that they author materials that focus on repetition and rehearsal. These recommendations of strategies could then be used to create the domain-dependent materials that will be used in the ITS. Additionally, in certain situations, such as when the material is largely question and survey based, the system could ask the author to enter their questions and then order them as appropriate based on characteristics of the questions (e.g., to harness the serial position effect if relevant), and in a specific order to ensure that material is rehearsed by the individual (e.g., to improve moving from short-term to long-term memory). This information that was entered could then be populated into a tutor, instead of the author needing to generate separate content. This authoring tool could also provide a means for authors to identify if they would like to use the learner's name throughout the tutoring (e.g., to elicit the self-reference effect), and provide a structure for the author to create questions that will use the names. In the case of GIFT, many of these features could either be built into existing authoring tools or as a separate tool to assist with course construction.

4 Conclusions

A great deal of research has been conducted in the field of cognitive psychology, and much of it is relevant to adaptive instruction as it highlights strategies for improving memory and learning. ITS authors can improve their tutors by being aware of different effects or principles that could help or hurt the memory retention that their learners have, and the creation of a potential authoring aid or authoring tool in an ITS system could assist them in creating content that is consistent with these principles.

Acknowledgements. The research described herein has been sponsored by the U.S. Army Research Laboratory. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Atkinson, R.C., Shiffrin, R.M.: Human memory: a proposed system and its control processes. *Psychol. Learn. Motiv.* **2**, 89–195 (1968)
2. Broadbent, D.E.: The role of auditory localization in attention and memory span. *J. Exp. Psychol.* **47**(3), 191 (1954)
3. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. *Cogn. Psychol.* **12**(1), 97–136 (1980)
4. Moray, N.: Attention in dichotic listening: affective cues and the influence of instructions. *Q. J. Exp. Psychol.* **11**(1), 56–60 (1959)
5. Sinatra, A.M., Sims, V.K., Najle, M.B., Chin, M.G.: An examination of the impact of synthetic speech on unattended recall in a dichotic listening task. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 55(1), pp. 1245–1249. Sage Publications, Los Angeles (2011)
6. Sottolare, R.A., Graesser, A., Hu, X., Holden, H. (eds.): Preface in design recommendations for intelligent tutoring systems: volume 1-learner modeling. US Army Research Laboratory (2013)
7. Sottolare, R.A., Brawner, K.W., Sinatra, A.M., Johnston, J.H.: An updated concept for a Generalized Intelligent Framework for Tutoring (GIFT) (2017). GIFTutoring.org
8. Wang-Costello, J., Goldberg, B., Tarr, R.W., Cintron, L.M., Jiang, H.: Creating an advanced pedagogical model to improve intelligent tutoring technologies. In: *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)* (2013)
9. Murdock Jr., B.B.: The serial position effect of free recall. *J. Exp. Psychol.* **64**(5), 482 (1962)
10. Baddeley, A.D., Hitch, G.: Working memory. *Psychol. Learn. Motiv.* **8**, 47–89 (1974)
11. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**(2), 81 (1956)
12. Baddeley, A.D.: Working memory. *Phil. Trans. R. Soc. Lond. B* **302**(1110), 311–324 (1983)
13. Strayer, D.L., Johnston, W.A.: Driven to distraction: dual-task studies of simulated driving and conversing on a cellular telephone. *Psychol. Sci.* **12**(6), 462–466 (2001)
14. Symons, C.S., Johnson, B.T.: The self-reference effect in memory: a meta-analysis. *Psychol. Bull.* **121**(3), 371 (1997)
15. Sinatra, A.M.: A Personalized GIFT: recommendations for authoring personalization in the generalized intelligent framework for tutoring. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) *AC 2015. LNCS (LNAI)*, vol. 9183, pp. 675–682. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20816-9_64
16. Sinatra, A.M., Sims, V.K., Sottolare, R.A.: The impact of need for cognition and self-reference on tutoring a deductive reasoning skill (No. ARL-TR-6961). Army Research Lab Aberdeen Proving Ground, MD (2014)
17. Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. *Science* **211**(4481), 453–458 (1981)
18. Arkes, H.R., Blumer, C.: The psychology of sunk cost. *Organ. Behav. Hum. Decis. Process.* **35**(1), 124–140 (1985)
19. Legge, E.L., Madan, C.R., Ng, E.T., Caplan, J.B.: Building a memory palace in minutes: equivalent memory performance using virtual versus conventional environments with the Method of Loci. *Acta Physiol.* **141**(3), 380–390 (2012)
20. Abernethy, E.M.: The effect of changed environmental conditions upon the results of college examinations. *J. Psychol.* **10**(2), 293–301 (1940)



Community Models to Enhance Adaptive Instruction

Robert Sottolare^(✉)

U.S. Army Research Laboratory, Orlando, FL, USA
robert.a.sottolare.civ@mail.mil

Abstract. This paper discusses the need and methods to develop community-based persona (learner models) to tie together key learner attributes and learning outcomes (e.g., knowledge acquisition) with the goal of facilitating the validation of adaptive instructional strategies and tactics. Adaptive instruction, sometimes referred to as differentiated instruction, is a learning experience tailored to the needs and preferences of each individual learner or team in which strategies (recommendations and plans for action) and tactics (actions by the tutor) are selected with the aim of optimizing learning, performance, retention, and the transfer of skills between the instructional environment (usually provided by an Intelligent Tutoring System or ITS) and the work or operational environment where the skills learned will be applied. Adaptive instructional systems (AISs) use human variability and other learner attributes along with instructional conditions to select appropriate strategies and tactics. This is usually accomplished through the use of machine learning techniques, but large amounts of data are needed to reinforce the learning of these algorithms over time. We propose a method to develop community models more quickly by enabling diverse groups to contribute the results of their experiments and training data in a common instructional domain to a cloud-based model that could be shared by various instructional applications.

Keywords: Adaptive Instructional Systems (AISs)
Intelligent Tutoring Systems (ITSs) · Learner modeling
Reinforcement learning

1 Introduction

Adaptive instructional systems (AISs) provide machine-based instruction through technologies like Intelligent Tutoring Systems (ITSs) which interact with learners and make decisions about interventions based on the needs and preferences of each individual learner [1]. These interventions are based on a model of that learner or team and the conditions in the instructional environment or application. A simple model of instruction includes instructional elements (Fig. 1) to be considered during the AIS authoring process [2].

In most ITSs, learning objectives are defined by the author and outline the terminal goals of the instruction. For example, a tutor that guides/adapts instruction for a marksmanship task might have a learning objective, “understand how to maintain a

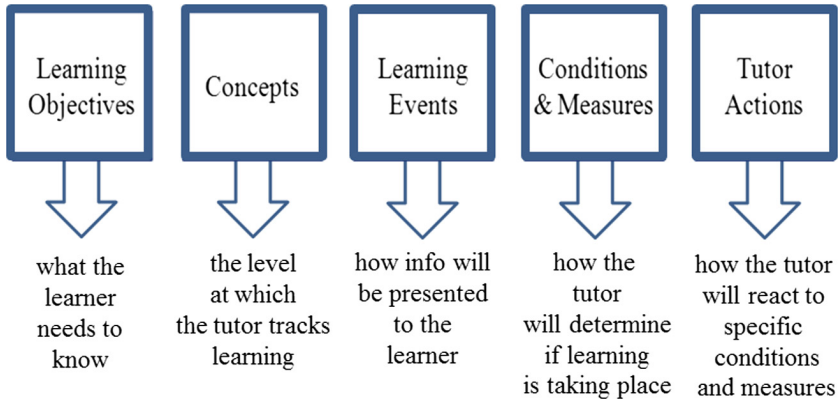


Fig. 1. Elements of an adaptive instructional model

steady position during weapons fire”. In a tutoring architecture, like the Generalized Intelligent Framework for Tutoring (GIFT) [3, 4], the objectives are set as concepts to be learned. This is usually the level at which the tutor tracks individual or team learning. Learning events represent content offered to the learner and focuses on how and what the learner experiences. Conditions and measures set standards against which the tutor determines the learner’s progress toward learning objectives. Finally, the tutor will react to changing learner attributes (e.g., change in competency level) and environmental/application conditions (e.g., place in the instruction) by intervening with the learner. This intervention could take the form of feedback, interaction with the learner (e.g., ask a question, prompt the learner for more information or provide a hint), or changes to the environment (e.g., increase the difficulty level of the task).

This model usually includes critical information about the learner and the instructional domain that informs a machine learning algorithm in the tutor and that algorithm is trained by consuming data involving both successful and unsuccessful decisions. Decision success/failure is based on their effect on learning outcomes which include: knowledge and skill development, retention, performance, and transfer of skills from instructional to operational (work) environments.

A basic adaptive instructional model (Fig. 2) involves learner actions and conditions, environmental conditions, instructional policies, and interactions (actions, observations, and assessments) capturing data between the tutor, the environment (sometimes referred to as the application), and the learner.

The instructional model and its policies can be improved over time through a class of machine learning techniques called reinforcement learning (RL) algorithms. RL algorithms are often used to improve the accuracy and reliability of adaptive instructional decisions. However, this method requires usually large amounts of data to develop optimally effective instructional policies that drive tutor strategies and tactics. To reduce the development time of instructional policies, we advocate a mechanism to collaboratively develop instructional models for a variety of task domains. Before we discuss community models, it is appropriate to review how reinforcement learning works in practice.

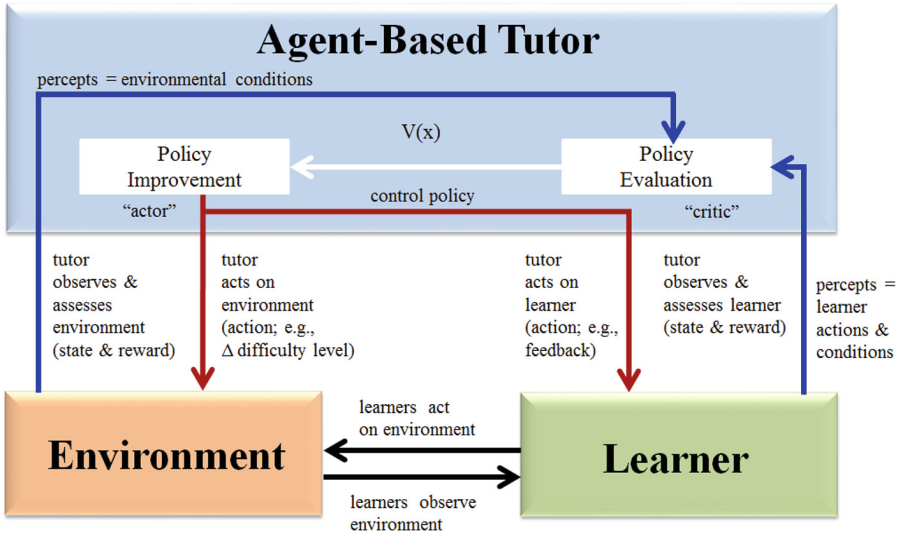


Fig. 2. Interaction within an instructional model

2 Reinforcement Learning for Learner and Instructional Models

Reinforcement learning, a type of machine learning, enables software agents to automatically determine an optimal action within a specific context in order to maximize its own performance over time [5]. Reward feedback is all that is required for a software agent to learn or reinforce the selection of optimal behavior(s). The process of selecting optimal behaviors over time and many examples is known as a Markov Decision Process (MDP) [6].

Conditional systems within AISs are used to determine tutor behavior or more specifically their decisions and interactions with the learner and the environment. These interactions are in the form of instructional strategies and tactics in systems like GIFT [3, 4]. AISs consider instructional strategies and tactics which are bounded by constraints posed by policies and these policies are often framed in terms of Markov Decision Processes (MDPs) [7] which seek to reinforce maximum outcomes or rewards over time [6, 8]. To select an optimal policy or modify it based on new information, the MDP considers the current state and the value of any actions which might transition to successor states with respect to a desired outcome.

According to Mitchell [9], the optimal action, a , for a given state, s , is any action that maximizes an immediate reward, $r(s, a)$ and the value, V , of the immediate successor state, s' . What does this mean for AISs? First, based on our model in Fig. 2, states are much more complex in AISs than in most systems. States must capture conditions of the learner (e.g., performance, affect, competence) and the instructional environment (e.g., concept under instruction, concept map (hierarchical relationships between learning objectives), and recent content presented) that affect the learning

experience. Actions, referred to as tactics in GIFT, are the set of instructional options available in the current state.

The reward function is tied directly to learning outcomes (e.g., knowledge and skill development, performance, retention, and/or transfer from instructional to operational or work environments) and the value is the anticipated performance in the next state. It is easy to see that the number of possible states can be very large in AISs and that the task of validating MDPs for all possible states could take a single organization a very long time. Hence the need for a process to divide the validation process into smaller discrete elements that can be processed by researchers in parallel, but to a similar standard.

The idea would be to have learner models grow quickly based on the individual and team instruction received by a learner in a variety of instructional environments in the learning landscape (e.g., formal education, training, reading, job-related tasks). Instructional models which might form the basis of widespread policies and strategies or tactics in specific domains could be evaluated across a variety of populations with emphasis on what works best for novices, moderates, and experts in any given domain. In this way it would be possible to generalize instructional techniques that work broadly and label them as domain-independent policies or strategies.

3 An Approach to the Development of a Community Model

Based on our goal to reduce the time to validate a complex instructional model, let's simplify our model for adaptive instruction by dividing it into its three essential elements: the learner model, the instructional environment, and the instructor or tutor. The learner model consists of the attitudes and behaviors along with the cognitive states of the learner. This model could also include long term information that highlights trends, habits, preferences, interests, values, and other data that could influence learning.

The instructional environment consists of learning objectives (LOs; also known as concepts), a concept map (a hierarchical relationship of concepts to be learned as shown in Fig. 3), a set of learning activities which include content and directions on how the learner will interact with the content, a set of measures to determine learning and performance, and a set of available tutor strategies and tactics to respond to various learner attitudes, behaviors and cognitive states.

Examining the concept map in Fig. 3, we see that it illustrates prerequisite relationships among nine concepts (A-I) with A-D showing the learner mastered those concepts. Concepts E, G, H, and I have not been attempted, but F has and the results show that the learner has not yet attained mastery of this concept. This type of map helps the tutor understand the relationship of concepts, what the learner knows and does not know, and provides context for machine learning algorithms to select appropriate instructional strategies.

The tutor also consists of a set policies that drive its behavior and interaction with the learner and the instructional environment. The goal is for the policies to be updated regularly as the tutor interacts with more and more learners and finds new highs to override previous best practices. According to Chi and Wylie [10], learner activities vary from least effective to most effective are: passive (receiving), active

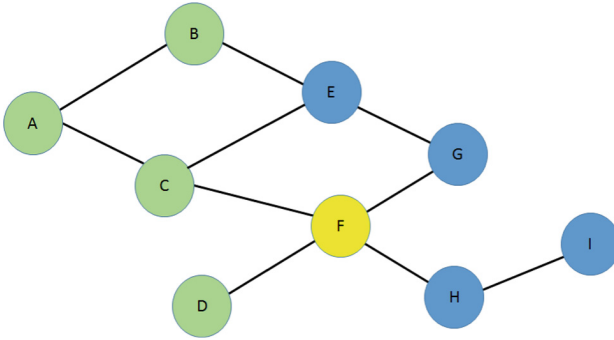


Fig. 3. Mastery policy in a concept map within an instructional model

(manipulating), constructive (generating), and interactive (dialoguing). As activities are selected and presented to the user, the tutor uses measures to assess progress toward learning and performance. For example, in a tutor that instructs learners to read, the tutor might engage the learner in a reflective dialogue (interactive activity) about a recently read passage to ascertain the learner’s comprehension of the concepts presented in that specific reading.

To further our approach, we might consider generalizing terms and measures (Table 1) in lieu of using specific measures. Reducing the number of discrete states also reduces the matrix for selecting the best possible response by the tutor to existing conditions.

4 Discussion

By coming to consensus on a common set of terms and defining their relationships in an ontology [11], we might realize the degree of interoperability needed to develop community-based models for AISs. However, we also realize that their complexity [12] and the lack of interoperability [13] between various AISs may slow the progress of developing these models. The good news is that current events have highlighted significant opportunities to capture and share the data needed to grow community learner and instructional models.

Recently, the Institute of Electrical and Electronics Engineers (IEEE) Learning Technologies Steering Committee (LTSC) approved a study group to examine opportunities for standards to promote interoperability and reuse with this class of technologies known as AISs [14]. The modularity of systems may play a large part in motivating organizations to share data and resulting models. If successful, the IEEE LTSC’s initiative will likely result in a high degree of sharing among AIS components, tools, methods, and data.

In 2017, under the auspices of the North Atlantic Treaty Organization’s Human Factors and Medicine Panel (NATO-HFM), a research task group examining

Table 1. Instructional model element descriptors

Instructional model elements	Independent variables	Dependent variables	Variables of interest	Nominal descriptor
Learner model				
Attitudes			X	Pos - Neg
Behaviors			X	Behavioral marker
Cognitive states			X	H-M-L workload
Instructional environment				
Concepts (learning objectives)			X	LO description
Concept map (including prerequisites)				
Learning activities (ICAP - interactive, constructive, active, and passive)			X	ICAP
Content & interaction				
Directions & support				
Measures of learning & performance (desired outcomes)		X		H-M-L
Instructional policies	X			Policy name
Tutor strategies & tactics				

technologies and opportunities to exploit Intelligent Tutoring Systems for adaptive instruction completed its task and recommended “the development of standard learner model attributes which include both domain-independent (e.g., demographics) and domain-dependent (e.g., domain competency, past performance and achievements) fields which are populated from a learner record store (LRS) or long-term learner model. This will promote standard methods to populate real-time models during ITS-based learning experiences and allow for common open learner modeling approaches and transfer of competency models from one tutor to another” [15]. If adopted, this recommendation may be an impetus in creating a large, diverse community from which data for community learner models could be harvested.

While not a recent phenomenon, the advent of the educational data mining repositories DataShop [16] and its successor, LearnSphere [17] provide mechanisms for contributing and consuming experimental data related to learners interacting with instructional systems like AIs. “LearnSphere integrates existing and new educational data and analysis repositories to offer the world’s largest learning analytics infrastructure with methods, linked data, and portal access to relevant resources” [18].

5 Next Steps

There are some significant barriers to developing community models to enhance adaptive instruction. The first challenge is not technical... The rights to data are complex. The second challenge is developing models of teams is much more complex than developing models of individuals [19].

The first challenge is attracting organizations to contribute regularly to a database of instruction for the benefit of the community. DataShop and LearnSphere have a long history of incentivizing the learning science community to contribute. They have overcome the barrier of who owns the individual data and the rights to how it might be used by stripping out personally identifiable information (PII) so that no data can be associated with a specific individual. These sanitized databases form a basis for sharing without violating any individual rights. LearnSphere also has selective access in which it allows the collecting organization to determine who is allowed rights to use the data.

However, some of the individual details we mentioned previously (behaviors, trends, habits, preferences, and values) that might be useful in tailoring instruction, may not be shared in experimental databases today, but are part of our internet DNA. Websites like Amazon and Google collect information about us regularly to tailor our internet experiences (e.g., shopping, entertainment) and recommendations. Dino Wilkinson, an international attorney at Norton Rose Fulbright reported that “under English law, there are no property rights in data as such – although this has not necessarily prevented individuals and businesses from treating data as property. Markets exist for buying or selling data and individuals regularly disclose their personal data in exchange for goods and services. However, the value in these cases is created through the right to sell or use the data in a certain way rather than a legal right of ownership” [20]. Ultimately, it could be individuals who own their data, manage access to it, and license it for use by others. Until then, scientists who collect data have to be sensitive to its use. This will slow, but not stop the progress to model individuals for use in adaptive instructional systems.

The next horizon for AIs is to be easily applied in both individual and team instructional domains. The second challenge, team modeling, has some twists and turns that make their development much more complex than the development of individual models. The first twist is common sense: team tasks are more complex to assess than individual tasks since they involve individual members engaged in interdependent roles and responsibilities in pursuit of a goal or a set of goals. This means a team of X members has X sources of data, and many sets of interactions which may be important to measure and assess during machine-based team tutoring.

Teams are involved in the process of teamwork and the learning and maintenance of team skills required by the team taskwork. Teamwork involves “coordination, cooperation, and communication among individuals to achieve a shared goal” [21]. Teamwork is largely a domain-independent process and includes the social skills (e.g., tact and trust) needed to function as an effective team. The interaction of teamwork on team taskwork is prevalent in the literature [22–25], and the antecedent attitudes, behaviors and cognition has been recently analyzed and defined in structural equation models derived from a major meta-analysis of the team and tutoring literature [26].

A large part of the complexity of modeling teams is embedded in the difficulty in acquiring and interpreting learner data and interaction data. Sensors, self-report mechanisms, external observations, and historical databases may be sources to inform team models, but they should be unobtrusive to avoid any negative impact on the learning process.

Collecting data is a challenge, but filtering that data to create information to inform tutor decisions is another challenge and a major contributor to the complexity of modeling teams. The development of accurate machine learning methods to select optimal tutor interventions to enhance team learning and performance is desirable and difficult. The sheer number of conditions present for individual learners on the team, the instructional environment, and the options available to the tutor are mind boggling, but perfect for machines.

As noted for individuals, team models, including models of instruction, may benefit from simple analysis of effectiveness where comparisons between models of teams, instruction, and domains are used to identify significant differences in learning, performance, retention, or transfer of learning from instruction to operations. Of course we need to build team tutors first [27], but eventually we need a playground to experiment and test. A testbed model used in GIFT [28] and based on Hanks et al. [29] has been used for evaluating adaptive instruction of individuals and might easily be extended to support team tutoring (Fig. 4).

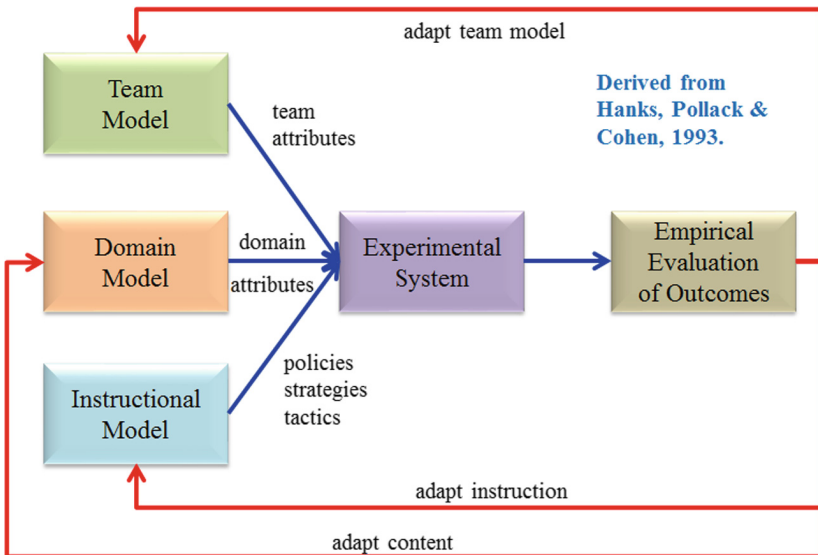


Fig. 4. Testbed for evaluating the effect of adaptive instructional of teams

In a modular system like GIFT, the team, instructional and domain models have common interfaces and message types that inform data shared between the models. This allows flexibility in swapping out or just changing the internal processes within

these models to allow experimentation with different policies, strategies, and tactics in varying conditions represented in learner/team models and the environment or application (e.g., problem set, simulation or webpage). For example, an experimenter could examine instructional strategies for low performing teams in game-based simulation scenarios.

Since GIFT is also very data-centric, experimenters are permitted change out one set of parameters for another. For example, an experimenter could decide to examine instructional interventions for the common three-tier performance model in GIFT (at, above or below expectations) for a more granular performance model with five or more levels.

Given the complexity of examining the large number of conditions represented in the learner/team, instructional, and domain models with associated content, we believe it will be significantly faster and easier to share the analysis and development workload through community modeling schema.

References

1. Sottolare, R.: A comprehensive review of design goals and emerging solutions for adaptive instructional systems. *Technol. Instr. Cogn. Learn. (TICL)* **11**, 5–38 (2018)
2. Sottolare, R.A.: Adaptive instruction for individual learners within the Generalized Intelligent Framework for Tutoring (GIFT). In: Schmorow, D.D., Fidopiastis, C.M. (eds.) *AC 2016. LNCS (LNAI)*, vol. 9744, pp. 90–96. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39952-2_10
3. Sottolare, R.A., Brawner, K.W., Goldberg, B.S., Holden, H.K.: The Generalized Intelligent Framework for Tutoring (GIFT). Concept paper released as part of GIFT software documentation. U.S. Army Research Laboratory – Human Research & Engineering Directorate (ARL-HRED), Orlando (2012). https://gifttutoring.org/attachments/152/GIFTDescription_0.pdf. Accessed
4. Sottolare, R., Brawner, K., Sinatra, A., Johnston, J.: An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). US Army Research Laboratory, Orlando, May 2017. <https://doi.org/10.13140/rg.2.2.12941.54244>
5. Singh, S.P.: Reinforcement learning algorithms for average-payoff Markovian decision processes. In: *AAAI*, vol. 94, pp. 700–705, October 1994
6. Puterman, M.L.: *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York (2014)
7. Cassandra, A.R.: *Exact and approximate algorithms for partially observable Markov decision processes* (1998)
8. Nye, B., Sottolare, R., Ragusa, C., Hoffman, M.: Defining instructional challenges, strategies, and tactics for adaptive intelligent tutoring systems. In: Sottolare, R., Graesser, A., Hu, X., Goldberg, B. (eds.) *Design Recommendations for Intelligent Tutoring Systems: Volume 2 - Instructional Management*. Army Research Laboratory, Orlando (2014). ISBN 978-0-9893923-2-7
9. Mitchell, T.M.: *Machine Learning*. McGraw-Hill, Boston (1997). ISBN 978-0-07-042807-2
10. Chi, M.T.H., Wylie, R.: The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ. Psychol.* **49**(4), 219–243 (2014). <https://doi.org/10.1080/00461520.2014.965823>

11. Bechhofer, S.: OWL: web ontology language. In: Liu, L., Özsu, M.T. (eds.) *Encyclopedia of Database Systems*, pp. 2008–2009. Springer, Boston (2009)
12. Sottolare, R., Ososky, S.: Defining complexity in the authoring process for adaptive instruction. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) *AC 2017. LNCS (LNAI)*, vol. 10285, pp. 237–249. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_17
13. Wegner, P.: Interoperability. *ACM Comput. Surv. (CSUR)* **28**(1), 285–287 (1996)
14. Sottolare, R., Brawner, K.: Exploring standardization opportunities by examining interaction between common adaptive instructional system components. In: *Proceedings of the First Adaptive Instructional Systems (AIS) Standards Workshop*, Orlando, Florida, March 2018
15. Sottolare, R.A.: Chapter 8 – Summary and Recommendations for the Exploitation of ITS Technologies in NATO. NATO Final Report of the Human Factors & Medicine Research Task Group (HFM-RTG-237), Assessment of Intelligent Tutoring System Technologies and Opportunities. NATO Science & Technology Organization (2018). <https://doi.org/10.14339/sto-tr-hfm-237>. ISBN 978-92-837-2091-1
16. Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: the PSLC DataShop. In: *Handbook of Educational Data Mining*, p. 43 (2010)
17. Stamper, J., Koedinger, K., Pavlik Jr., P.I., Rose, C., Liu, R., Eagle, M., Veeramachaneni, K.: Educational data analysis using LearnSphere workshop. In: *Proceedings of the EDM 2016 Workshops and Tutorials co-located with the 9th International Conference on Educational Data Mining*, Raleigh, NC. Workshop (2016)
18. LearnSphere: A community data infrastructure to support learning improvement online (2018) <http://learnsphere.org/index.html>. Accessed
19. Sottolare, R., Holden, H., Brawner, K., Goldberg, B.: Challenges and emerging concepts in the development of adaptive, computer-based tutoring systems for team training. In: *Proceedings of the Interservice/Industry Training Simulation & Education Conference*, Orlando, Florida, December 2011
20. Howell, D.: The cloud conundrum: who actually owns your data? TechRadar (2014). <https://www.techradar.com/news/internet/cloud-services/the-cloud-conundrum-who-actually-owns-your-data-1260464>. Accessed
21. Salas, E.: *Team Training Essentials: A Research-Based Guide*. Routledge, London (2015)
22. LePine, J.A., Piccolo, R.F., Jackson, C.L., Mathieu, J.E., Saul, J.R.: A meta-analysis of teamwork processes: tests of a multidimensional model and relationships with team effectiveness criteria. *Pers. Psychol.* **61**(2), 273–307 (2008)
23. Lim, B.C., Klein, K.J.: Team mental models and team performance: a field study of the effects of team mental model similarity and accuracy. *J. Organ. Behav.* **27**(4), 403–418 (2006)
24. Ellis, A.P., Bell, B.S., Ployhart, R.E., Hollenbeck, J.R., Ilgen, D.R.: An evaluation of generic teamwork skills training with action teams: effects on cognitive and skill-based outcomes. *Pers. Psychol.* **58**(3), 641–672 (2005)
25. Crawford, E.R., Lepine, J.A.: A configural theory of team processes: accounting for the structure of taskwork and teamwork. *Acad. Manag. Rev.* **38**(1), 32–34 (2013)
26. Sottolare, R.A., Burke, C.S., Salas, E., Sinatra, A.M., Johnston, J.H., Gilbert, S.B.: Designing adaptive instruction for teams: a meta-analysis. *Int. J. Artif. Intell. Educ.* **28**(2), 225–264 (2017)
27. Bonner, D., Walton, J., Dorneich, M.C., Gilbert, S.B., Sottolare, R.A.: The development of a testbed to assess an intelligent tutoring system for teams. In: *Workshop on Developing a Generalized Intelligent Framework for Tutoring (GIFT): Informing Design Through a Community of Practice*. Seventeenth International Conference on Artificial Intelligence in Education (AIED 2015), p. 9 (2015)

28. Sottilare, R., Goldberg, B., Brawner, K., Holden, H.: A modular framework to support the authoring and assessment of adaptive computer-based tutoring systems (CBTS). In: Proceedings of the Interservice/Industry Training Simulation & Education Conference, Orlando, Florida, December 2012
29. Hanks, S., Pollack, M.E., Cohen, P.R.: Benchmarks, test beds, controlled experimentation, and the design of agent architectures. *AI Mag.* **14**(4), 17 (1993)



Biocybernetic Adaptation Strategies: Machine Awareness of Human Engagement for Improved Operational Performance

Chad Stephens¹(✉), Frédéric Dehais², Raphaëlle N. Roy²,
Angela Harrivel¹, Mary Carolyn Last³, Kellie Kennedy¹,
and Alan Pope¹

¹ NASA Langley Research Center, Hampton, VA 23681, USA
{chad.l.stephens, angela.r.harrivel, kellie.d.kennedy,
alan.t.pope}@nasa.gov

² Institut Supérieur de l'Aéronautique et de l'Espace (ISAE),
Université de Toulouse, Toulouse, France
{frederic.dehais, raphaelle.roy}@isae-superaero.fr

³ Analytical Mechanics Associates, Hampton, VA 23666, USA
mary.c.last@nasa.gov

Abstract. Human operators interacting with machines or computers continually adapt to the needs of the system ideally resulting in optimal performance. In some cases, however, deteriorated performance is an outcome. Adaptation to the situation is a strength expected of the human operator which is often accomplished by the human through self-regulation of mental state. Adaptation is at the core of the human operator's activity, and research has demonstrated that the implementation of a feedback loop can enhance this natural skill to improve training and human/machine interaction. Biocybernetic adaptation involves a "loop upon a loop," which may be visualized as a superimposed loop which senses a physiological signal and influences the operator's task at some point. Biocybernetic adaptation in, for example, physiologically adaptive automation employs the "steering" sense of "cybernetic," and serves a transitory adaptive purpose – to better serve the human operator by more fully representing their responses to the system. The adaptation process usually makes use of an assessment of transient cognitive state to steer a functional aspect of a system that is external to the operator's physiology from which the state assessment is derived. Therefore, the objective of this paper is to detail the structure of biocybernetic systems regarding the level of engagement of interest for adaptive systems, their processing pipeline, and the adaptation strategies employed for training purposes, in an effort to pave the way towards machine awareness of human state for self-regulation and improved operational performance.

Keywords: Biocybernetic adaptation · Adaptive training · Engagement

1 Introduction

Human operators generally face a complex, dynamic and uncertain environment under time pressure. The occurrence of unexpected events (e.g., critical failure) requires flexibility and cognitive regulation policies to meet task demand (Sperandio 1978). Various strategies may be employed by humans to achieve adaptation. In a study of the physiological effects of a kinetically adaptive environment, Jager et al. (2017) describe the reciprocal relationship between adaptive humans and such environments. Schwarz and Fuchs (2017) point out that “humans are adaptive systems themselves”, that is, they are able to mitigate critical user states by applying self-regulation strategies. They cite as examples “investing more effort if task demands increase or drinking coffee to combat fatigue”. Additionally, the system itself can be made to adapt to the human.

Some examples of the integration of simultaneous human and system adaptation are aimed at psychophysiological goal achievement, not necessarily at the immediate achievement of optimal performance. Biocybernetic adaptation has been employed as a self-regulation training method for application in clinical and sports settings (Pope et al. 2014). In these technologies, the adaptation approach involves physiological signals modulating some aspects of the training tasks in such a way as to reward trainees for approaching a target signal. These physiological self-regulation training technologies are designed to improve adherence to a training regimen by delivering the training through engaging, motivating, and entertaining experiences. The processing employed in these technologies is minimal to enable real-time feedback. Likewise, the decision rules are usually simple, e.g., modulating a single task element based upon signal level. For instance, some consequences in a digital game or simulation reward the user for achieving a psychophysiological goal by diminishing an undesirable effect in a game (analogous to negative reinforcement). Other consequences reward the user for achieving a psychophysiological goal by producing a desirable effect (analogous to positive reinforcement) such as additional scoring opportunities. That is, some modulation effects enable superimposed disadvantages in a digital game or simulation to be reduced by progression toward a psychophysiological goal, whereas others enable advantages to be effected by progression toward a psychophysiological goal.

Schmorrow proposes a system that adapts to a trainee’s level in the context of flight training: “Imagine an aviation recruit experiencing a simulator that is tailored to the trainee at the most fundamental neurophysiological level. Imagine that this simulator’s integrated helmet and sensor suite are hooked up to a ‘black box’ that modifies the simulated flight exercise based on a real-time assessment of the student pilot’s cognitive state, using information collected by the sensor suite.” (Schmorrow 2005). Similarly, the task modulation concept embodied in the self-regulation training technology based on biocybernetic adaptation may be adapted for use in task simulators. The simulator embodiment of the closed-loop modulation concept, Stress Counter-response Training (Palsson and Pope 1999), integrates physiological self-regulation training into the practice of mission-relevant tasks. Stress Counter-response Training is based upon the concept of instrument functionality feedback which ties the functionality of a simulator to the requirement to maintain the

physiological equanimity suited for optimal cognitive and motor performance under emergency events in an airplane cockpit.

In these technologies, the physiological modulation method is tailored to the overall game or simulation task, but without regard for changes in the task context or other situational factors. Fuchs and Schwarz (2017) identify this as a “hard-coded” adaptation strategy, where the system triggers a predetermined adaptation strategy. As will be shown, even more complex adaptation strategies have considerations in common with the simple self-regulation training strategy.

2 Biocybernetic Loop Implementation for Adaptive Systems

A first and important step for biocybernetic adaptation is to determine what temporal and magnitude changes in physiological signals reflect operator or trainee state changes that warrant mitigation (Fairclough and Gilleade 2013). Indeed, one important concern with the implementation of such assisting systems is to succeed in providing assistance in a timely and appropriate manner (Parasuraman et al. 1999). Spurious triggering of the assistance system may have negative consequences on human operators (Parasuraman et al. 1997). Therefore, an approach is to target mental states that are (1) relevant predictors of human performance and (2) that can be robustly identified via behavioral and neurophysiological measures. Mental states of interest are discussed, followed by a description of the biocybernetic adaptation pipeline.

Traditionally, most of the research has focused on mental workload-based biocybernetic adaptation. However, the usability of the mental workload construct remains limited. Although theoretically and practically interesting, it remains ill-defined (Mandrick et al. 2016), providing a non-specific and generic index rather like a thermometer. Moreover, mental workload should not be viewed as the result of an external demand applied on an individual passively adapting to it, but rather as an active process that depends on the human operator’s level of engagement. For instance, a highly demanding situation will not necessarily induce high workload if an individual does not engage to achieve. Several reasons may account for this lack of engagement such as excessive task difficulty (Durantin et al. 2014), repetitive and boring tasks (Durantin et al. 2015) and cognitive fatigue (Hopstaken et al. 2015). Conversely, over-engagement in a non-priority and non-demanding task could induce high workload (e.g., interacting with the entertainment system or texting while driving) and jeopardize safety (Lee 2014; Dehais et al. 2012). Thus, human cognitive performance has to be considered the byproduct of the level of task demand by the level of task engagement. Interestingly, the concept of engagement is related to a triad of attentional states: attentional disengagement, attentional over-engagement, and attentional in-engagement. Also, the study of engagement is richer than the concept of workload: this concept accounts for neurophysiological and behavioral phenomena and it can be characterized with portable measurement tools (Verdiere et al. 2018). For example, a biocybernetic system was designed to mitigate task disengagement due to automation by triggering changes in task mode based on the fluctuations of an engagement index constructed as a ratio of EEG band powers (Scerbo et al. 2000). Derivation of the engagement index was based on the proposition that the closed-loop paradigm that

represents the adaptive configuration in which physiological indices are to have a steering role can also serve as a prior validation test bed for the indices themselves (Pope et al. 1995).

Firstly, attentional disengagement occurs when task demand is too low leading to episodes of mind wandering (Durantin et al. 2015) or when task demand exceeds mental capacity. In these two extreme situations, human operators generally drop the primary task to focus on automatic secondary tasks. These two states are characterized by the disengagement of the executive network, underpinned by the deactivation of the dorsolateral prefrontal cortex (Durantin et al. 2014; Harrivel et al. 2013). Secondly, attentional over-engagement, also referred to as attentional tunneling (Wickens 2005) and “channelized attention” (Harrivel et al. 2016), is defined as “the allocation of attention to a particular channel of information, diagnostic hypothesis or task goal, for a duration that is longer than optimal, given the expected cost of neglecting events on other channels, failing to consider other hypotheses, or failing to perform other tasks”. Some authors postulate that this impaired attentional state results from a disengagement deficit of the orientation network underpinned by the thalamus (LaBerge et al. 1992). Whereas the assessment of such brain structure remains difficult to be performed in operational context - it requires the use of fMRI – some studies have disclosed that attentional over-engagement is associated with an attentional shrinking and long fixation time (Dehais et al. 2011). Recently, the EEG engagement index proposed by Pope et al. (1995) was shown to be sensitive to episodes of over-engagement leading to inattentional deafness to auditory alarm under real-flight settings (Dehais et al. 2014).

Lastly, recent work has shown the existence of an attentional in-engagement state whereby human operators are unable to engage their attention to process relevant information when facing critical situations. One could describe this state as “panic mode” in a vernacular fashion. This state, that is the exact opposite of attentional tunneling, is explained in terms of impaired thalamus tonic mode to maintain focused attention. This state of “attentional confusion” or “attentional entropy” is associated with high saccadic activity and absence of long fixations (Dehais et al. 2015).

Another interesting approach could be to identify the dynamic model of such features. Tools derived from the linear algebra and control communities can be applied to perform an approximation of the neurophysiological features model that could be explored to monitor the engagement of an operator. The method provides a smooth interpolation of all the data points enabling the extraction of frequency features that reveal fluctuations in engagement with growing time-on-task (Poussot-Vassal et al. 2017). Alternatively, the use of large-scale EEG connectivity is a relevant approach not only to detect but also to predict future performance and fluctuation of engagement (Senoussi et al. 2017).

The implementation of the biocybernetic adaptation pipeline mostly consists of the classical steps of a Brain-Computer Interface, that is to say a signal acquisition step (e.g., EEG), a preprocessing step that generally deals with artifacts (e.g., eye blinks) and better conditions the signal, a feature extraction step (e.g., extraction of the average power in specific frequency bands), a machine learning step (e.g., a classification step), and lastly an adaptation step (Roy and Frey 2016). This last step can consist of providing the estimated mental state to the system’s decisional unit. The decisional unit system allows the loop to be closed. This is done by implementing a decisional unit

driven by a policy resulting from the resolution of a (Partially Observable) Markov Decision Process ((PO)MDP) that takes into consideration uncertainties on actions, partial observable states (i.e., mental states) or potentially non-deterministic behavior of the human operator (Gateau et al. 2016; Drougard et al. 2017). Eventually, a last step is to design a catalogue of adaptive solutions to mitigate decline in performance and improve human performance. These solutions are presented in the next section.

3 Successful Implementation of Adaptive Solutions

Self-regulation training can be deliberate as described earlier or could occur inadvertently as a result of an operator's exposure to an adaptive system. Technology in the field of self-regulation training has commonly taken into account the fact that the physiological self-regulation behavior and skill of the trainee changes as training progresses. These systems have incorporated algorithms that respond to momentary, transient changes in physiological signals in real time, as well as longer time course changes that reflect a trainee's emerging ability to voluntarily control physiological parameters. The momentary changes are displayed as information and reward feedback for learning of self-regulation skill, while the longer time course measurements are assessed to guide the setting of higher and higher self-regulation performance goals.

An early example is an electromyographic biofeedback training system that implemented a shaping procedure by adjusting the gain of the feedback loop after each interval of training based on a trainee's success at lowering EMG levels (Pope and Gersten 1977). This system employed a fixed strategy by which task characteristics are adapted to the individual. A training strategy implies a set of assertions relating strategy characteristics and their effects on training progress. In a more advanced implementation, a data base of these assertions could be updated on-line and the training system would be self-improving. In effect, the system would evaluate the results of mini-experiments with various strategy versions within a session and modify the strategy accordingly. O'Shea and Sleeman (1973) developed this hierarchical framework in the context of adaptive teaching systems.

Similarly, physiologically adaptive systems will need to be designed to respond appropriately not only to transient changes and spontaneous drifts in operator state due to developing conditions such as fatigue, but also to conditioning of physiological changes as a result of an operator's extended exposure to information feedback about their physiological state. Accordingly, an adaptive implementation that took into consideration the operator "training" effect of its information feedback employed a continually updated model of the operator analogous to the "template of average performance" in the "symbiotic cockpit" (Reising and Moss 1985). Techniques developed for adapting a brain-computer interface classifier to adjust for possible features drift could be applied to address this type of consequence (Vidaurre et al. 2011). Configuration of an adaptive system that takes into consideration these long-term and short-term processes is depicted in Fig. 1.

An additional strategy is the deliberate exercise of self-regulation skill acquired as a result of self-regulation training. Prinzel et al. (2002) demonstrated that participants given feedback of the accuracy of their estimates of engagement levels, across a

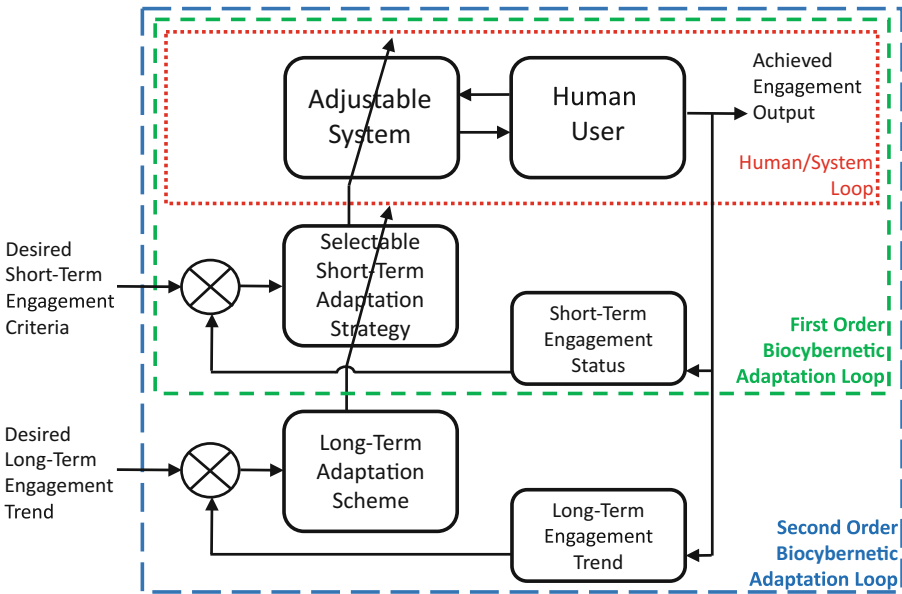


Fig. 1. Configuration of an adaptive system for managing human user engagement. The Desired Short-Term Engagement Criteria and Long Term Engagement Trends vary according to context (e.g., phase of flight). The Adjustable System has adjustable parameters, such as automation level. The Adjustable System is adjusted by the currently invoked Short-Term Adaptation Strategy and is driven by input from the human user. The Selectable Short-Term Adaptation Strategy has a catalog of selectable strategies, each one of which is in effect for a selectable duration. The currently invoked Strategy is selected by the Long-Term Adaptation Scheme and is driven by the discrepancy between the desired engagement criteria and actual engagement status. The Long-Term Adaptation Scheme is driven by the discrepancy between the desired and actual engagement trend. The Achieved Engagement Output is the level of an engagement index derived from physiological and behavioral measures and is driven by the effects of the Adjustable System on the Human User.

partitioning of the range of an EEG-based engagement index into six subranges, were able to achieve “a 70% level of correct identifications.” Further, while interacting with an adaptive automation system, “Participants in the self-regulation condition were better able to maintain their task engagement level within a narrower range of task modes, thereby reducing the need for task mode changes. The effect of this was an increase in task performance as well as a decrease in reported workload.” Prinzel et al. (2002) further comment, “The neurofeedback provided during training may have allowed these participants to better manage their cognitive resources and thereby regulate their engagement state, allowing them to better respond to a change in automation mode. The results of this study support other research that has shown that physiological self-regulation could enhance the cognitive resource management skills of pilots and complement the benefits of adaptive automation.” An outcome of self-regulation training accomplished with an adaptive system is improved cognitive state management skill which is effectively meta-awareness on the part of the trainee.

This effect demonstrates an observing system as defined in second order cybernetics (von Foerster 1995).

In addition to clinical and sports applications, biocybernetic adaptation as self-regulation training is applied in a third area, the aircrew training context (Stephens et al. 2017). In this application, the instructor-trainee interaction is influenced, closing the loop on a broader time scale. Here the adaptation involves an attention management training approach to complement the usual observations of airline training instructor pilots by informing them, in the training context, of the occurrence of attention-related human performance limiting states (AHPLS) experienced by their trainees. Classifier models are trained to recognize trainee state during simulated flight scenarios based on patterns of the physiological signals measured during benchmark tasks (Harrivel et al. 2016). Machine learning models' real time determinations of the cognitive states induced by the scenario tasks are displayed as gauges embedded in a mosaic of windows that also displays real time images of the scenario tasks that the trainee is performing (e.g., scene camera, simulator displays, animation of simulator controls), and this mosaic¹ is video recorded (Harrivel et al. 2017). The loop is closed when their state information is conveyed to the trainee as part of each session debrief. This approach involving trainee-trainer interaction leverages the effective bio-social influences on learning specified by Kamiya (Strehl 2014). Like the adaptive automation application, the adaptation strategy here takes into consideration contextual parameters such as the instructor's discretion regarding the appropriateness of conveying particular state information to the trainee.

This psychophysiological-based AHPLS detection and mitigation system is modeled after the Hypoxia Familiarization Training (HFT) employed in aviation. The focus of HFT is on recognizing symptoms of hypoxia and taking steps to recover from the hypoxia being experienced. Similarly, recognition and recovery from AHPLS is intended to improve self-monitoring of and response to one's own attentional performance, maintaining more effective states and managing attention. Such meta-awareness results from this form of self-regulation training intended to develop attention management skill. If deployed in ground-based commercial aviation training contexts, the intent is to mitigate potential in-flight loss of airplane state awareness (ASA) and thus reduce aviation accidents and incidents.

Biocybernetic adaptation can be applied within autonomous systems to imbue further intelligence into the systems about the humans involved in operations. In a potential adaptive automation application, the cognitive state of the operator of a semi-autonomous vehicle would be tracked by the vehicle system. The system uses the cognitive state information to judge the operator's ability to take back control of the system in critical or noncritical hand-off instances².

¹ This concept is captured in a non-provisional patent application: Stephens et al. (2017, patent pending) "System and Method for Training of State-Classifiers." [NASA Case No.: LAR-18996-1].

² This concept is captured in a non-provisional patent application: Harrivel et al. (2017, patent pending) "System and Method for Human Operator and Machine Integration." [NASA Case No.: LAR-19051-1].

4 Conclusion

The implementations of specific adaptations described in this paper represent actual systems designed for improving human/machine interaction and furthermore enabling human operators to improve self-regulation skills. The adaptation strategies described herein include combinations of technological advances in the areas of neuroscience and psychophysiology designed for specific contexts including clinical, aviation, and sports. The example implementation systems instantiate concepts and enable practical and empirical testing to evaluate adaptation strategies. Adaptation management issues were discussed including dynamic selection and configuration of adaptations. Development of adaptation strategies can create further questions for consideration such as how to handle possible side effects on the human operator caused by setting up a biocybernetic loop. This and other questions require empirical results to be sufficiently addressed. Ongoing research efforts at NASA and the Institut Supérieur de l'Aéronautique et de l'Espace (ISAE) seek to apply adaptation strategies to answer these questions and reveal further questions with the ultimate goals of improved safety and efficiency in aerospace operations.

References

- Dehais, F., Causse, M., Tremblay, S.: Mitigation of conflicts with automation: use of cognitive countermeasures. *Hum. Factors* **53**(5), 448–460 (2011)
- Dehais, F., Causse, M., Vachon, F., Régis, N., Menant, E., Tremblay, S.: Failure to detect critical auditory alerts in the cockpit: evidence for inattentive deafness. *Hum. Factors* **56**(4), 631–644 (2014)
- Dehais, F., Causse, M., Vachon, F., Tremblay, S.: Cognitive conflict in human automation interactions: a psychophysiological study. *Appl. Ergon.* **43**(3), 588–595 (2012)
- Dehais, F., Peysakhovich, V., Scannella, S., Fongue, J., Gateau, T.: Automation surprise in aviation: real-time solutions. In: *Proceedings of the 33rd Annual ACM conference on Human Factors in Computing Systems*, 2525–2534 (2015)
- Drougard, N., Chanel, C.P.C., Roy, R.N., Dehais, F.: Mixed-initiative mission planning considering human operator state estimation based on physiological sensors. In: *IROS17, 9th Workshop on Planning, Perception and Navigation for Intelligent Vehicles* (2017)
- Durantini, G., Gagnon, J.F., Tremblay, S., Dehais, F.: Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behav. Brain Res.* **259**, 16–23 (2014)
- Durantini, G., Dehais, F., Delorme, A.: Characterization of mind wandering using fNIRS. *Front. Syst. Neurosci.* **9**, 45 (2015). <https://doi.org/10.3389/fnsys.2015.00045>
- Fairclough, S., Gilleade, K.: Capturing user engagement via psychophysiology: measures and mechanisms for biocybernetic adaptation. *Int. J. Auton. Adapt. Commun. Syst.* **6**(1), 63–79 (2013)
- Fuchs, S., Schwarz, J.: Towards a dynamic selection and configuration of adaptation strategies in augmented cognition. In: Schmorow, D., Fidopiastis, C. (eds.) *AC 2017, Part II. LNCS*, vol. 10285, pp. 101–115. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_7

- Gateau, T., Chanel, C.P.C., Le, M.-H., Dehais, F.: Considering human's non-deterministic behavior and his availability state when designing a collaborative human-robots system. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4391–4397. IEEE (2016)
- Harrivel, A.R., Weissman, D.H., Noll, D.C., Peltier, S.J.: Monitoring attentional state with fNIRS. *Front. Hum. Neurosci.* **7**, 861 (2013)
- Harrivel, A., Liles, C., Stephens, C., Ellis, K., Prinzel, L., Pope, A.: Psychophysiological sensing and state classification for attention management in commercial aviation. In: American Institute of Aeronautics and Astronautics, SciTech 2016, San Diego, California (2016)
- Harrivel, A., Stephens, C., Milletich, R., Heinich, C., Last, M.C., Napoli, N., Abraham, N., Prinzel, L., Motter, M., Pope, A.: Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing. In: American Institute of Aeronautics and Astronautics, SciTech 2017, Grapevine, Texas (2017)
- Hopstaken, J.F., Linden, D., Bakker, A.B., Kompier, M.A.: A multifaceted investigation of the link between mental fatigue and task disengagement. *Psychophysiology* **52**(3), 305–315 (2015)
- Jager, N., Schnädelbach, H., Hale, J., Kirk, D., Glover, K.: Reciprocal control in adaptive environments. *Interact. Comput.* **29**(4), 512–529 (2017)
- LaBerge, D., Carter, M., Brown, V.: A network simulation of thalamic circuit operations in selective attention. *Neural Comput.* **4**, 318–331 (1992)
- Lee, J.D.: Dynamics of driver distraction: the process of engaging and disengaging. *Ann. Adv. Automot. Med.* **58**, 24 (2014)
- Mandrick, K., Chua, Z., Causse, M., Perrey, S., Dehais, F.: Why a comprehensive understanding of mental workload through the measurement of neurovascular coupling is a key issue for neuroergonomics? *Front. Hum. Neurosci.* **10** (2016). <https://doi.org/10.3389/fnhum.2016.00250>
- O'Shea, T., Sleeman, D.: A design for an adaptive self improving teaching system. In: Rose, J. (ed.) *Advances in Cybernetics and Systems*, vol. 3. Gordon & Breach, London (1973)
- Palsson, O.S., Pope, A.T.: Stress counterresponse training of pilots via instrument functionality feedback. on symposium: new methods in biofeedback delivery: NASA innovations from aerospace to inner space. In: *Proceedings of the 1999 Applied Psychophysiology (AAPB) Meeting*, 10, April 1999, Vancouver, Canada (1999)
- Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* **39**, 230–253 (1997)
- Parasuraman, R., Mouloua, M., Hilburn, B.: Adaptive aiding and adaptive task allocation enhance human-machine interaction. In: Scerbo, M.W., Mouloua, M. (eds.) *Automation Technology and Human Performance: Current Research and Trends*, pp. 119–123. Erlbaum, Mahwah (1999)
- Pope, A.T., Gersten, C.D.: Computer automation of biofeedback training. *Behav. Res. Methods Instrum.* **9**, 164–168 (1977)
- Pope, A.T., Bogart, E.H., Bartolome, D.S.: Biocybernetic system validates index of operator engagement in automated task. *Biol. Psychol.* **40**, 187–195 (1995)
- Pope, A.T., Stephens, C.L., Gilleade, K.: Biocybernetic adaptation as biofeedback training method. In: Fairclough, S., Gilleade, K. (eds.) *Advances in Physiological Computing*. HCIS, pp. 91–115. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6392-3_5
- Poussot-Vassal, C., Roy, R.N., Bovo, A., Gateau, T., Dehais, F., Chanel, C.P.C.: A loewner-based approach for the approximation of engagement-related neurophysiological features. Presented at the 20th The International Federation of Automatic Control (IFAC) World Congress, Toulouse, France, July 2017 (2017)

- Prinzel, L.J., Pope, A.T., Freeman, F.G.: Physiological Self-regulation and adaptive automation. *Int. J. Aviat. Psychol.* **12**(2), 179–196 (2002)
- Reising, J.M., Moss, R.W.: 2010: the symbiotic cockpit. In: Proceedings of the National Aerospace and Electronics Conference, Dayton, OH, vol. 2, 20–24 May 1985, pp. 1050–1054 (1985)
- Roy, R.N., Frey, J.: Neurophysiological markers for passive brain–computer interfaces. In: Clerc, M., Bougrain, L., Lotte, F. (eds.) *Brain-Computer Interfaces 1: Foundations and Methods*. Wiley, Hoboken (2016)
- Scerbo, M.W., Freeman, F.G., Mikulka, P.J.: A biocybernetic system for adaptive automation. In: Backs, R.W., Boucsein, W. (eds.) *Engineering Psychophysiology: Issues and Applications*, pp. 241–253. Lawrence Erlbaum, Mahwah (2000)
- Schmorrow, D.D.: Aviation Training: A Future Avenue. *Avionics Magazine*, October 2005
- Schwarz, J., Fuchs, S.: Multidimensional Real-Time Assessment of User State and Performance to Trigger Dynamic System Adaptation. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) *AC 2017, Part I. LNCS (LNAI)*, vol. 10284, pp. 383–398. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58628-1_30
- Senoussi, M., Verdier, K.J., Bovo, A., Chanel, C.P.C., Dehais, F., Roy, R.N.: Pre-stimulus antero-posterior EEG connectivity predicts performance in a UAV monitoring task. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1167–1172. IEEE, October 2017
- Sperandio, J.C.: The regulation of working methods as a function of work-load among air traffic controllers. *Ergonomics* **21**(3), 195–202 (1978)
- Stephens, C., Prinzel, L., Harrivel, A., Comstock, R., Abraham, N., Pope, A., Wilkerson, J., Kiggins, D.: Crew state monitoring and line-oriented flight training for attention management. In: Proceedings of the 19th International Symposium on Aviation Psychology (ISAP), 8–11 May 2017, Dayton, OH (2017)
- Strehl, U.: What learning theories can teach us in designing neurofeedback treatments. *Front. Hum. Neurosci.* **8** (2014). <https://doi.org/10.3389/fnhum.2014.00894>
- Verdière, K.J., Roy, R.N., Dehais, F.: Detecting pilot’s engagement using fNIRS connectivity features in an automated vs manual landing scenario. *Front. Hum. Neurosci.* **12**, 6 (2018)
- Vidaurre, C., Sannelli, C., Müller, K.-R., Blankertz, B.: Machine-learning-based coadaptive calibration for brain-computer interfaces. *Neural Comput.* **23**(3), 791–816 (2011)
- von Foerster, H. (ed.): *Cybernetics of cybernetics: Or, the control of control and the communication of communication*, 2nd edn. Future Systems, Minneapolis (1995)
- Wickens, C.D.: *Attentional Tunneling and Task Management*. Technical report, AHFD-05-01/NASA-05-10, NASA Ames Research Center, Moffett Field CA (2005)

Brain Sensors and Measures for Operational Environments



Do Not Disturb: Psychophysiological Correlates of Boredom, Flow and Frustration During VR Gaming

Klaas Bombeke¹✉, Aranka Van Dongen¹, Wouter Durnez¹,
Alessandra Anzolin², Hannes Almgren², Anissa All¹, Jan Van Looy¹,
Lieven De Marez¹, Daniele Marinazzo²,
and Elena Patricia Núñez Castellar^{1,2}

¹ Imec-mict-UGent, Ghent University, Ghent, Belgium
klaas.bombeke@ugent.be

² Department of Data Analysis, Ghent University, Ghent, Belgium

Abstract. Since the technology behind virtual reality (VR) is evolving rapidly and the number of VR applications is growing every year, research on the user's experience of being in the virtual environment itself and the methodologies to measure these experiences becomes highly important. In this study, we apply the methodology of measuring attentional allocation by means of a dual-task paradigm to the topic of VR gaming. The idea is to ask participants to react to oddball sounds (secondary task), pulling attention away from the primary task (the game). The behavioral (reaction time and accuracy) and neural response (P300 component) to these oddball sounds then tells us something about indirect attentional allocation to the game and possibly the experience of flow. In order to check the latter, we created experiences of boredom, flow and frustration by manipulating the mechanics of the game. In addition, we were interested in other psychophysiological correlates like brain oscillations and average heart rate and whether these differed between gaming with a regular and a VR set-up. Although we were not able to accurately induce feelings of boredom, flow and frustration and could not replicate previous studies showing increased reaction times for oddball sounds during flow, we did find a decreased P300 and more high-frequency brain oscillations in VR compared to regular gaming (indicating more attention to the game). Together, this suggests that psychophysiological measures are promising tools to quantify attentional allocation in VR, but more research is needed to clarify whether and how this translates to flow.

Keywords: Virtual reality · Gaming · User experience · Flow
EEG · ERP · P300

1 Introduction

After more than two decades of technological improvements, virtual reality has evolved to an artificially created environment in which the user feels actually present [1]. Presence can be defined as the sense of “being there” in which the body behaves as if it is part of the illusion [2]. To a great extent, this feeling of presence is accomplished by

using high-definition computer-generated graphics and stereoscopic 3D images that are presented on a head-mounted display (HMD), which moves along with the user's head. Virtual reality found its way to everyday entertainment, but also to industrial, educational and medical settings [3, 4]. For example, doctors are trained to perform surgery using virtual reality, patients suffering from Parkinson's disease learn to deal with their decreased postural control [5], and psychologists try to overcome the client's social phobias with virtual environments built for exposure therapy [6]. Since the technology behind virtual reality (VR) and head-mounted display (HMD) devices is changing rapidly and the number of VR applications is growing every year, research on the user's experience of being in the virtual environment itself and the methodologies to measure these experiences has become highly important too. However, it is important to clearly define what is meant with "experience", which is a rather vague and broad concept. In the context of gaming, experience is defined as 'an ensemble made up of the player's sensations, thoughts, feelings, actions and meaning-making in a gameplay setting' [7]. Moreover, considering that it points to an interaction between game, gamer and context, it is an agent-dependent and highly subjective concept [8]. Consequently, it is suggested that it is better to look at game experience as an underlying mechanism that make games motivating and fun.

Therefore, an important framework that is often referred to when studying the user's experience in games is 'flow theory' [9]. Flow can be defined as a state of intense attentional focus, pleasurable feelings, and emotional rewards when engaged in a certain activity or task [e.g., 10–12]. Examples of activities that can produce a flow experience are dancing, making music, exercising, but also playing video games [13, 14]. Moreover, flow is often associated with a distorted perception of time and a loss of self-consciousness. It is also a highly individual-specific experience because activities that are perceived as enjoyable by some can be dreary for others. The key principle of flow is that it is determined by the sweet spot between skill and challenge [10, 11]. When a person's skills are highly developed but the challenge is low, he or she becomes bored. In contrast, when skills are underdeveloped but the challenge is high, a feeling of frustration and anxiety might arise. Flow occurs when skills and challenge are perfectly aligned, leading to satisfaction and happiness. In order to create an optimal gaming experience related to flow, clear goals, feedback, sense of control and an appropriate balance between challenge and skills should be integrated [9, 15]. More specifically, the overriding goal of the game should be made clear from the beginning and intermediate goals should be presented at appropriate times [16, 17]. Feedback can consist of several elements: progress towards the goals, immediate feedback on in-game actions and the ability to know your status or score in the game at any given time [15]. Challenge is considered the most important aspect of a good game design. Difficulty levels of a video game should be variable to meet all players at the correct level of challenge. To create this balance between challenge and skill level, games typically start out with a beginner's level which gradually increases in difficulty as the player's skills progress [17, 18].

Interestingly, an important characteristic of flow during gaming is the high level of attentional focus. When playing a game, the player has to allocate attention to relevant stimuli (e.g. opponents, targets) but ignore irrelevant stimuli (e.g. a tree in the background). In the flow state, these irrelevant stimuli coming from surroundings are usually more ignored and less consciously processed compared to a state without

feelings of flow [19]. This led to an interesting line of research using dual-task paradigms to measure flow in an indirect way [e.g., 20, 21]. The idea is that in addition to a primary task (i.e. playing the video game), a secondary task is implemented in which stimuli are presented that will attract the player's attention. A good candidate for the secondary task is the classic oddball paradigm, originating from the field of cognitive psychology. In this paradigm, a series of auditory tones are presented at more or less the same pace. The crucial manipulation is that in one out of ten tones (by approximation), participants hear a deviant 'oddball' tone that is presented with a different tonal frequency and the participants are asked to press a button as soon as they hear it. The idea is that this oddball stimulus pulls attention away from the primary task (i.e. playing the game) and can therefore be an indirect marker of the amount of attention allocated to this primary task. Flow would be associated with increased attentional focus, so the response to the oddball tone should be slower. Interestingly, an oddball tone also elicits the so-called P300 component. This P300 event-related potential component reflects an increase in electrical activity on the midline posterior scalp surface around 300 ms after the presentation of the tone and has been related to cognitive surprise or stimulus categorization [e.g., 22]. Similar to the behavioral response and based on the assumption that there is a limited set of attentional resources [23], it follows that when people are actively engaged in the primary task (or video game), fewer attentional resources will be available for the processing of the auditory tones, which will be reflected by a decreased P300 response. Furthermore, because of this close relationship between flow and attention, Weber and colleagues approached flow from a neurocognitive point of view and proposed their 'Synchronization theory of flow', which can be directly applied to a media context [12]. In this theory, flow is considered as a synchronization phenomenon of different attentional and reward networks in the brain. Decades of research on brain oscillations has shown that when brain regions oscillate with similar frequencies (or "rhythms"), less energy is consumed. In other words, neural synchronization is energetically cheap [24]. This can explain why flow is rarely associated with feelings of exhaustion or burdensomeness, although the task or activity can be highly challenging [12]. Evidence for this theory was found in multiple experiments [20, 25, e.g., 26–28]. Weber et al. [25], for example, showed increased functional connectivity among attentional networks with decreasing distraction to stimuli in a secondary task.

However, in this study, our primary goal was to extend the methodology of the dual task-paradigm to the domain of virtual reality. This is especially interesting, because although questionnaires and interviews are quite common when measuring gaming experience [29], it can be argued that they are likely to disrupt the immersive experience in VR. For example, when a participant is playing a game in a virtual environment without any other avatars in the room, the voice of an experimenter or a pop-up message asking about his experience will likely decrease his or her feelings of presence. In this regard, the use of electrophysiological measurements like EEG (brain activity), GSR (galvanic skin response) and ECG (heart rate) can be very informative. These kinds of psychophysiological measures are not only less likely to disrupt the immersive experience in VR, they also reflect unconscious cognitive and affective processes that are impossible to measure with questionnaires or via interviews. However, a challenge where the field currently has to deal with is the fact that

psychophysiological signals often become hard to interpret when studying real-life behavior like playing a game in VR. There are two main reasons for this. First, since it is often impossible to impose rigorous experimental control [30], the signal-to-noise ratio is too low to obtain reliable measurements. The solution for this is to increase the duration (and hence quantity) of the measurements, in order to obtain more data so the noise will be averaged out. A second reason relates to the fact that the visual stimulation in a virtual environment can be so complex and intense that it becomes hard to distinguish between different cognitive and affective processes [21]. On top of that, processes with increased activation (e.g. more attention to targets) and processes with decreased activation (e.g. less attention to background) might camouflage each other in the physiological signals, so they will not be picked up in the analysis. We wanted to check whether using the dual-task paradigm to indirectly measure attention allocated to the primary task would be a solution for these issues in the context of VR.

In order to compare different attentional states determined by the balance between the challenges and skills of an individual, we modified the game mechanics of a popular shooter game so that an experience of boredom, flow and frustration was created. In addition, we let participants experience all three conditions both with a regular PC gaming set-up (desktop computer connected to television screen; from now on referred to as 'PC') and a VR set-up (desktop computer connected to HTC Vive HMD; from now on referred to as 'VR'). While participants played the shooter game with their right hand, they had to respond to oddball sounds with the left hand. Concretely, eighteen participants played the same game with subtle adaptations for eight minutes each in six different versions (fully counterbalanced within-subjects design; VR/PC x boredom/frustration/flow) while their EEG was recorded, heart rate was monitored, subjective flow experience was questioned and in-game performance was logged.

With respect to our predictions, we first wanted to make sure whether the participants did indeed experience the boredom, flow and frustration condition in the way we intended it, by analyzing their self-reports (i.e. flow questionnaire). Second, since previous research has shown that the response to the oddball sounds was delayed and that more detection errors were made when participants experienced flow compared to boredom and frustration [21], our main research question was whether we could replicate this effect and whether this effect would be larger for VR. The latter can be expected because immersion is higher in VR. Similarly, we hypothesized that in VR, compared to regular gaming on a PC set-up, participants would show a smaller P300 component, indicating increased attentional allocation to the game. Furthermore, we expected this effect to be driven by an experience of flow, compared to an experience of boredom or frustration. Third, with respect to brain oscillations, we expected to find changes in alpha power (8–13 Hz), where increased alpha power would be an indication of boredom and decreased alpha power an indication of attentional focus. In addition, increases in high-frequency power (beta and gamma) would also reflect more attentional allocation. However, we were not sure whether the auditory stimuli in the oddball paradigm would have an effect on the oscillatory activity related to the secondary task, making it hard to make exact predictions. Finally, we also included electrophysiological heart rate measurements (ECG) in order to measure arousal. Since increased flow and increased immersion in VR would likely lead to increases in autonomic arousal, we expected this to be reflected in the ECG signal.

2 Method

2.1 Participants

For this study, 18 participants were recruited through online sampling. There were more men (83%; $M = 23$ years old; $SD = 2.1$ years old) than women (17%; $M = 25.67$ years old; $SD = 2.52$ years old) and almost all of them (94.44%) were highly educated or still attending university. With respect to gaming experience, 61% classified themselves as casual gamers, 33% identified themselves as experts and 6% had never gamed before. In addition, 44% had already played the commercial version of presented game. All included subjects participated on a voluntary basis and signed an informed consent with ethical approval from the universal ethical committee.

2.2 Stimulus Material

Primary Task. As a primary task, subjects were asked to play a custom-made, first-person shooter game based on the commercial success “Counter-Strike: Global Offensive” (CS: GO). We chose this type of game because it is straightforward to play and has clear goals and immediate feedback, which are important prerequisites for the experience of flow [31]. The subject could start the game by triggering the slide door with the inscription “Start”. In each condition, the player started in a practice room with three targets on the wall. Next, there were five different rooms with targets they had to shoot (see Fig. 1). These targets were cardboard cut-outs of enemies that had to be shot twice, in order to avoid random shooting and accidental striking. Players could only proceed to the next room when every target in a room was hit. Immediate feedback was provided (“Good job”) when shooting all targets in a room and the subject could keep track of his/her progress and munition on a scoreboard. After five rooms, players arrived in a sixth room with a final scoreboard and a portal to the first level again, allowing them to start all over.



Fig. 1. Basic floorplan of the game with target placement (upper left) and decoration (bottom left) and screenshot of gameplay depicted in two versions of the map (upper and bottom right).

The three conditions of interest (boredom, flow and frustration) were operationalized by manipulating two different features of the game. First, the conditions differed in the speed at which the shooting targets moved. In the boredom condition, targets were stationary, making it rather easy to shoot them. In the flow condition, targets moved at a low speed from left to right and from back to forth. In the frustration condition, targets moved with high speed. Importantly, only in the flow condition, the speed of targets was adapted from room to room. By implementing this, we ensured that only in the flow condition subjects would make progress based on their skill level. A second manipulation concerned the amount of ammunition granted in each condition. In the boredom conditions, players had 12 bullets per target, which was plenty. In flow, we used a scheme for distributing ammunition throughout the levels, adapting it to the skills of the player. Players were only granted two bullets per target in the frustration condition, giving them no ammunition to spare. Because participants went through six gaming sessions of eight minutes each, two versions of the same map were alternately used, keeping elements like target placement, direction in which targets move, the amount of obstructions and the type of decoration as equal as possible (see Fig. 1).

Secondary Task. The goal of the secondary task in this experiment was to draw attention away of the primary task. We chose for the oddball paradigm that was also used in the study of Núñez Castellar et al. [21] and Debener and colleagues [32]. In this oddball paradigm, participants had to listen to auditory stimuli that were presented with random intervals of 960, 1060, 1160, 1260 and 1360 ms. Participants were instructed to react as fast and accurate as possible to the oddball sounds by means of a response box right below the keyboard that was used for the primary task. Importantly, the experimenter emphasized the importance of performing well on both the primary (i.e. the game) and secondary task. During each gaming session of eight minutes (six in total; two devices x three conditions), 320 sounds were presented. Each session consisted of 80% standard tones, 10% oddball sounds, and 10% novel sounds. The standard and oddball sounds were two sinusoids (350 Hz and 650 Hz) with a mean duration of 339 ms, whereas the 96 unique novelty sounds [33] had a random frequency with a mean duration of 338 ms. To avoid confounds, the low (350 Hz) and high (650 Hz) sinusoids were counterbalanced across participants alternating as standard or rare (oddball) sounds. For this study, only the standard and oddball sounds were analyzed.

2.3 Procedure and Design

In this study, we chose for a within-subjects design. This meant that for each participant, we could compare the three experience conditions (boredom, flow and frustration) for two different devices (2D and VR), without confounding effects of differential baselines (signal-to-noise ratio of the measurements or level of skills related to gaming performance) that we would have had with a between-subjects design. Concretely, participants had to play the three versions of the game (the boredom, flow and frustration version) both in virtual reality with the content presented on the HMD and in 3D (not stereoscopic) on a large television screen. To deal with the challenge of having delicate scalp measurements and a large head-mounted display with straps on top of the

head, the game was programmed in a way that the viewing direction could also be controlled with the mouse and keyboard. By doing so, minimal head movement was necessary to play the game, while a certain level of immersion in the virtual environment was reached. With respect to technical specifications, an Alienware gaming PC, a 46-in. Phillips television screen and a HTC Vive HMD were used. This HMD offered a resolution of 2160×1200 (with 1080×1200 per eye), global lighting and AMOLED-displays of 90 Hz. The task was played with a standard keyboard and mouse. The game itself was programmed in Notepad++ in the object-oriented open source programming language Squirrel. In order to run the game, CS: GO was opened through Steam, an online gaming platform developed by Valve. To convert CS: GO to the virtual reality domain, VorpX, a 3D-driver for virtual reality headsets with full head tracking support, was used.

To avoid training effects, the order of conditions was counterbalanced across participants, whereas device and condition were additionally counterbalanced within participants. Participants were tested individually and were seated at approximately 1 m of the screen behind a table (Fig. 2). While mounting the EEG electrodes, participants were asked to fill out the informed consent. Next, participants were asked to focus on a fixation cross for six minutes, while alternating between no blinking and being relaxed for one minute. Afterwards, standardized instructions were given to the subjects and they got the chance to practice navigation and shooting in the game environment. In between the different runs, participants filled out the Flow Questionnaire (FQ; [34]), which we used to measure the subjective experience of flow during gaming.



Fig. 2. Illustration of the experimental set-up with the participants playing the game in the virtual reality condition. The computer on the left provides the oddball task, whereas the laptop in the middle is used to record the EEG and the gaming computer with flat screen and the HTC Vive on the right display the game.

2.4 Electrophysiological Recordings and Preprocessing

EEG data was collected with a Biosemi ActiveTwo system (Biosemi, Amsterdam, Netherlands) using 64 Ag-AgCl scalp electrodes positioned according to the standard international 10–20 system. Because of the HMD, it was impossible to attach VEOG

and HEOG electrodes, but we did measure the ECG signal with three additional external electrodes (attached to the left and right collarbone and the lower left ribcage). EEG signals were recorded with a sampling rate of 2048 Hz and pre-processed with a bandpass filter of 0.01–30 Hz. Data was processed and analyzed using EEGLAB [35] and ERPLAB [36]. The quality of the continuous EEG was manually examined and large episodes of random noise were deleted. Nevertheless, all datasets were eventually included in the current study. Electrode P2 showed excessive high frequency noise in most datasets and was therefore deleted and interpolated. The signal was re-referenced offline to frontal electrode FPz, but results were quite similar with AFz or the average of all electrodes as reference. In a next step, epochs, time-locked to the onset of the auditory stimulus (oddball or standard), were extracted with a time window of –200 to 2000 ms. In order to exclude remaining artefacts in the data, epoch rejection was applied, deleting epochs (equally distributed across conditions) containing activity below and above a threshold of $-70 \mu\text{V}$ and $70 \mu\text{V}$, respectively. Because we were mainly interested in central posterior regions, the artefact rejection was only applied to channels CPz, Pz and POz.

After calculating the average across epochs per condition, a grand average across subjects was computed. Whereas waveforms showing the different electrodes and topographies were based on the grand average, latency and amplitude measures were calculated per subject and condition. With respect to the P300 component, all plots and statistics were based on the central midline electrodes Cz, Pz and POz [37, 38]. We chose these electrodes because the P300 is most often measured in parietal and central regions and less often in frontal electrodes (although a difference can be made between the P3a and P3b, which we will not cover in this study, but see Polich [37] for an overview).

3 Results

3.1 Questionnaires

Flow Questionnaire (FQ). The flow questionnaire as used by Sherry and colleagues [34] consists of 12 items measuring three subscales: easiness (similar to boredom), flow and difficultness (similar to frustration). Participants completed this questionnaire after each gaming session, allowing us to validate whether feelings of boredom, flow and frustration were indeed experienced in the respective conditions. A repeated measures ANOVA with factors device (PC vs. VR), condition (boredom, flow, frustration) and subscale (boredom, flow, frustration) was conducted on the scores. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of subscale, $\chi^2(2) = 23.18, p < .001$ and the interaction between condition and subscale, $\chi^2(9) = 65.23, p < .001$. Therefore, for these effects, Greenhouse-Geisser corrected tests are reported. The main effect of device was marginally significant, $F(1, 17) = 3.66, p = .07, r = .42$, whereas the main effects of condition and subscale were

highly significant, $F(2, 34) = 11.69, p < .001, r = .64$ and $F(1.13, 19.26) = 38.26, p < .001, r = .83$, respectively. The interaction between device and condition was non-significant, $F(2, 34) = .63, p = .53, r = .6$, but the interaction between device and subscale did show a significant effect, $F(2,34) = 5.47, p < .01, r = .49$. Both the interaction between condition and subscale and the three-way interaction between device, condition and subscale showed a tendency toward significance, $F(1.81, 30.80) = 3.05, p = .07, r = .39$, and $F(4, 68) = 2.42, p = .06, r = .35$.

Taken together, the statistical analysis does not show convincing differences in flow between the different gaming sessions. Indeed, when looking at the FQ-scores, participants primarily had the feeling that the games were easy to play. In the PC condition, the score on the subscale ‘easy’ was even higher than the score on subscale ‘flow’, $t(17) = 2.43, p = .03$, indicating that the operationalization of flow might have been suboptimal. However, when comparing the subjective experience of flow between VR and PC gaming, participants reported significantly more flow during VR gaming compared to PC gaming, $t(17) = 3.46, p < .01$ (Table 1).

Table 1. Subjectively reported experience of flow, as measured with the Flow Questionnaire. The bold numbers indicate the condition that should have had the highest score based on the experimental manipulation of the gaming experience.

Scale			Easy	Flow	Difficult
Flow Questionnaire	VR	Boredom	21.89 (7.22)	13.17 (5.29)	4.39 (1.97)
	VR	Flow	18.67 (8.53)	16.83 (6.73)	6.17 (2.28)
	VR	Frustration	21.44 (7.08)	14.78 (7.07)	9.5 (4.24)
	PC	Boredom	23.28 (6.03)	11.11 (5.16)	4.44 (1.72)
	PC	Flow	20.72 (8.11)	13.33 (5.71)	5.17 (1.65)
	PC	Frustration	21.22 (8.66)	14.22 (6.33)	9.06 (3.32)

3.2 Behavior

A two-way repeated measures ANOVA was conducted to compare the effect of device and condition on the mean reaction times measured in the conditions boredom, flow and frustration with both PC and VR. There was no significant main effect of device, $F(1, 17) = .02, p = .905, r = 0$, nor was there a main effect of condition, $F(16, 2) = .514, p = .603, r = .17$. In addition, no interaction effect was found between device and condition, $F(2, 16) = 1.51, p = .239, r = .28$. A non-parametric Friedman test of differences among repeated measures was conducted for the error rate and rendered a Chi-square value of 3.84, which is non-significant ($p > .5$). To detect possible differences between error rates in the conditions, a post-hoc test was administered for every feasible pair. A Wilcoxon Signed-Ranks Test indicated no significant differences in error rates between conditions and devices (all $ps > .5$) (Table 2).

Table 2. Means and standard deviations for reaction times (RT in seconds) and error rates (ER) related to detecting oddball sounds.

Condition	RT		ER	
	M	SD	M	SD
VR Boredom	.746	.088	2.00%	2.27%
VR Flow	.745	.100	2.27%	2.08%
VR Frustration	.738	.081	5.07%	12.15%
PC Boredom	.725	.079	2.05%	1.95%
PC Flow	.741	.100	2.25%	2.53%
PC Frustration	.759	.115	5.07%	12.98%

3.3 P300

In order to make the P300 as comparable as possible across conditions in terms of baseline, standard trials were subtracted from oddball trials. For the statistical analysis, we took the mean amplitude at three posterior midline locations (Cz, Pz and POz) between 500 and 1500 ms after the onset of oddball or target sound. This allowed us to statistically validate any differences between the conditions on different topographical locations of the scalp surface. A repeated-measures ANOVA with factors electrode (Cz, Pz, POz), device (PC, VR) and condition (Boredom, Flow, Frustration) was performed on the data. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of electrode, $\chi^2(2) = 19.38$, $p < .001$, the interaction between electrode and condition, $\chi^2(9) = 47.7$, $p < .001$, the interaction between electrode and device, $\chi^2(2) = 13.27$, $p = .001$, and the interaction between electrode, condition and device, $\chi^2(9) = 31.52$, $p < .001$. Therefore, for these effects, Greenhouse-Geisser corrected tests are reported.

First, the main effect of electrode was highly significant, $F(1.14, 17.15) = 9.05$, $p < .01$, $r = .61$, whereas the main effects of condition and device were not, $F(2, 30) = .24$, $r = .13$ and $F(1, 15) = .34$, $p = .57$, $r = .15$, respectively. Second, the two-way interactions between factors electrode and condition and factors condition and device were also not significant, $F(1.89, 28.42) = 1.02$, $p = .37$, $r = .25$ and $F(2, 30) = 1.85$, $p = .18$, $r = .33$, respectively, just like the two-way interaction between electrode and device, $F(1.24, 18.61) = 3.11$, $p = .09$, $r = .13$. Finally, the three-way interaction between electrode, device and condition was also non-significant, $F(2.1, 31.49) = .23$, $p = .80$, $r = .12$.

Because the two-way interaction between factors electrode and device unexpectedly showed a small tendency towards significance, we decided to do the comparison across conditions for each electrode separately. Results did not change for electrodes Pz and POz, but we now did find a significant main effect of device for measurements taken at electrode Cz, $F(1, 17) = 6.71$, $p = .02$, $r = .53$. As predicted, the P300 response decreased when participants were playing in VR ($M = 3.27 \mu\text{V}$, $SD = 2.59 \mu\text{V}$) compared to playing on the PC ($M = 1.31 \mu\text{V}$, $SD = 2.49 \mu\text{V}$; see Fig. 3). The main effect of condition was also significant, $F(2, 34) = 3.54$, $p = .04$, $r = .42$, whereas the interaction between device and condition was not, $F(2, 34) = .15$, $p = .86$, $r = .09$.

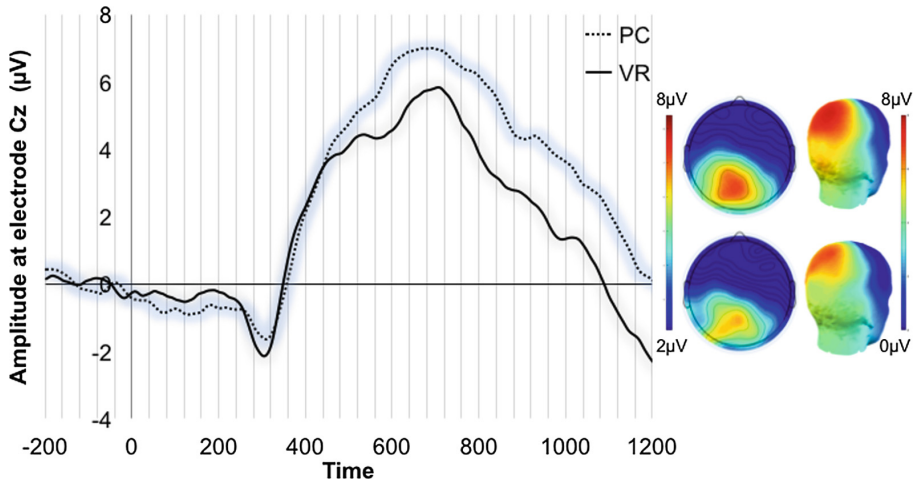


Fig. 3. The mean P300 component or the neural counterpart of detecting an oddball sound. The left side of the figure indicates that the amplitude of the P300 response was decreased when gaming in VR. The right side shows the topography and location where the P300 was maximal.

Focusing on electrode Cz, follow-up paired-samples *t*-tests gave some indication that this main effect of device was driven by the ‘flow’ condition, showing a marginally significant difference in P300 amplitude between 2D and VR, $t(17) = 2.05$, $p = .06$ (see Fig. 4), whereas the difference between 2D and VR for the boredom and frustration condition was not significant, $t(17) = .94$, $p = .36$ and $t(17) = 1.09$, $p = .29$.

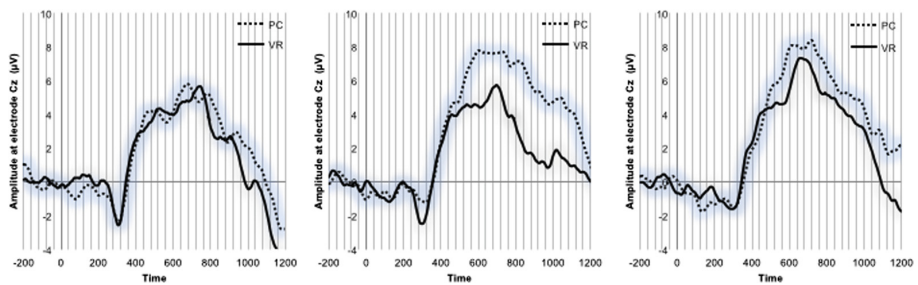


Fig. 4. The mean P300 component, split out for the different conditions (from left to right: boredom, flow and frustration). Although not significant, the plot suggests that the difference between regular and VR gaming is driven by the flow condition.

However, these findings should be interpreted with caution, since the topographical map in Fig. 3 makes it clear that the P300 component is more centered around Pz than around Cz, indicating that the significant main effect of device at Cz might not only reflect the P300 component.

When we correlated the mean P300 component across all conditions with the mean in-game performance score (average number of target hits), we found a highly significant negative correlation ($r = -.60$, $p < .01$, see Fig. 5). This means that when participants performed well, their P300 response was decreased and they were less distracted by the oddball sounds. All other correlations (corrected for multiple testing), did not reach significance (all $ps > .5$).

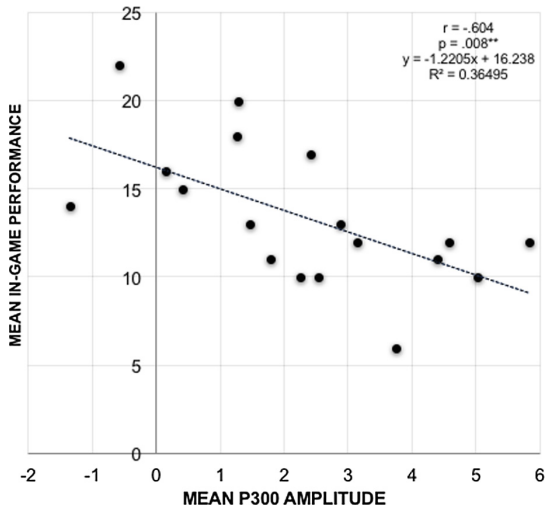


Fig. 5. Correlation between the mean in-game performance (number of targets hit) and the size of the P300 component.

3.4 Brain Oscillations

A repeated-measures ANOVA with factors frequency band (theta, alpha, low-beta, mid-beta, high-beta and gamma), device (PC and VR) and condition (boredom, blow, frustration) was performed on the power spectral density of gamma band oscillations, measured and collapsed across Cz, Pz and POz. Unsurprisingly, the main effect of frequency band was highly significant, $F(2.34, 39.73) = 35.11$, $p < .001$, $r = .82$. The main effects of device and condition were not significant, $F(1, 17) = .38$, $p = .55$, $r = .15$, and $F(1.05, 17.77) = .94$, $p = .35$, $r = .23$, respectively. With respect to the two-way interactions, we found a highly significant interaction between frequency band and device, $F(2.17, 36.85) = 9.58$, $p < .001$, $r = .60$. The interactions between frequency band and condition and between device and condition did not reach significance, $F(2.21, 37.53) = 2.58$, $p = .08$, $r = .36$, and $F(1.04, 17.61) = 1.14$, $p = .30$, $r = .25$, respectively. Finally, the three-way interaction between frequency band, device and condition was also not significant, $F(2.76, 46.87) = 1.34$, $p = .27$, $r = .27$. Because of the highly significant interaction between frequency band and device, we did some additional post-hoc analyses. Whereas the difference between PC and VR gaming was not significant for oscillatory power in the theta, alpha, low-beta and

mid-beta frequency ranges (all $ps > .1$), there was a significant increase for high-beta and gamma power spectral density in VR gaming compared to regular 2D gaming, $t(17) = 2.23, p = .04$ and $t(17) = 2.39, p = .03$, respectively (see Fig. 6).

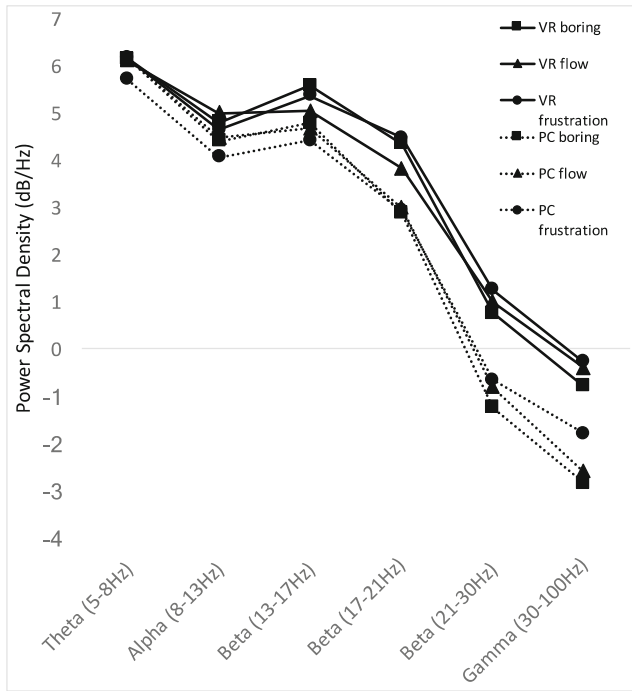


Fig. 6. Brain oscillations. The figure indicates that only high-frequency oscillatory activity was significantly increased when gaming in VR. There was no interaction with condition (boredom, flow and frustration)

Regarding the engagement index (beta power/alpha power), a repeated-measures ANOVA with factors device (PC and VR) and condition (Boredom, Flow, Frustration) was performed. The main effect of device was not significant, $F(1, 17) = 1.15, p = .30, r = .82$, just like the main effect of condition, $F(2, 34) = 1.04, p = .36, r = .78$. Also the interaction between device and condition did not reach significance, $F(2, 34) = .95, p = .40, r = .75$. Therefore, we did not explore this engagement index in further detail.

3.5 ECG

A repeated-measures ANOVA with factors device (PC and VR) and condition (boredom, flow, frustration) was performed on the average heart rate data (measured in beats per minute). The main effect of device was not significant, $F(1, 17) = .84, p = .37, r = .22$. There was no difference in average heart rate when participants were playing in VR ($M = 80.63$ bpm, $SD = 8.08$ bpm) compared to playing on the PC ($M = 78.69$ bpm,

$SD = 13.04$ bpm; see Fig. 7). The main effect of condition also lacked significance, $F(2, 34) = .06, p = .94, r = .06$, just like the interaction between device and condition, $F(2, 34) = .01, p = .99, r = .03$.

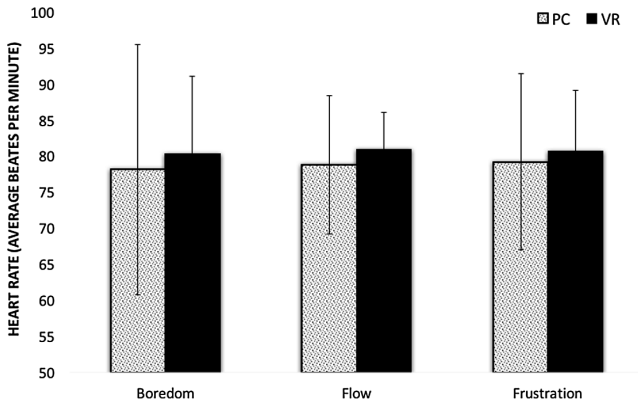


Fig. 7. Average heart rate, measured in beats per minute, based on the ECG signal.

4 Discussion

In this study, we expanded the work of Weber and Huskey [20] and Núñez Castellar et al. [21] and used their innovative dual-task approach to measure attentional allocation while gaming under conditions of boredom, flow or frustration. Importantly, we made a well-controlled comparison between regular PC and VR gaming in order to find out whether the dual-task approach could also be extended to more immersive technologies. The dual-task approach consisted of letting participants perform a primary task (shooter game) while asking them to respond to oddball sounds in a secondary task (oddball task). Every participant played six games in total (VR/PC x boredom/flow/frustration) and each game lasted eight minutes on average. Our first goal was to validate whether participants did indeed experience feelings of boredom, flow and frustration like we intended them to do. A second goal was to replicate the finding that during flow, participants were slower and less accurate to respond to oddball sounds than during boredom or frustration [e.g., 21] and to find out whether this effect would interact with the type of device (PC or VR). A third goal was to extend behavioral markers of attentional allocation during gaming to psychophysiological markers like the P300 ERP component, brain oscillations and the ECG signal.

Unfortunately, participants did not subjectively reported more feelings of flow in the flow condition compared to the boredom and frustration condition during both PC and VR gaming. Their predominant feeling was that of boredom in all conditions, indicating that the game might have been too easy to play. However, when comparing feelings of flow between PC and VR gaming, significantly more flow was experienced when gaming in VR. Hence, although our design was more ecologically valid (closer to real-life) than that of previous research (different game levels in each condition instead

of repetitions of levels in boredom and frustration condition, e.g. Núñez Castellar et al. [21]), these results indicate that the way in which we created feelings of boredom, flow and frustration still needs some adaptations and tweaking in order to get the experience right. Besides game mechanics, another reason why we failed to create experiences of boredom, flow and frustration might be related to how proficient participants were at playing games. Both experienced and non-experienced gamers participated, introducing potential confounds like having experience with being in a VR environment or not. This suboptimal experience of flow was also reflected in the behavioral data, for which we did not find significant effects of condition, device or their interaction. In contrast to Núñez Castellar et al. [21], reaction times to the oddball tones were surprisingly similar (around 742 ms) in all six conditions and we did not even observe a difference between PC and VR gaming. Given the results of the subjective and behavioral data, one could argue that it does not make sense to interpret the associated psychophysiological signals because the assumption that participants experienced boredom, flow and frustration in the respective conditions was not met. However, research in the past has often showed large dissociations between subjective and behavioral outcomes on the one hand and more unconscious, implicit physiological outcomes on the other hand [39] and the interpretation of these findings will at least convey interesting information for follow-up research. Therefore, we still think it is valuable to report our findings.

With respect to the neural processing of the oddball sound of the secondary task, we expected the largest decrease in P300 amplitude when participants experienced flow in VR. As expected, we observed a clear P300 response starting around 400 ms after the oddball sound, reaching its optimum at its typical posterior midline location around electrode Pz [e.g., 38]. Although we did not find significant main effects of condition or device, we did observe that the P300 was significantly decreased when playing in VR compared to playing with a regular PC set-up. However, this effect was driven by activity around the more anterior region (Cz), which did not entirely fit with the topographical location of the component. Furthermore, there was some indication that the decreased P300 for VR compared to PC was driven by the flow condition, but this effect should be interpreted with caution (marginally significant effects). Interestingly, we also observed a highly significant negative correlation between the average in-game performance score of a participant and the mean amplitude of the P300 (both collapsed across the six games). This basically means that the better a participant performed, the less he or she was distracted by the oddball sounds. Assuming this correlation can be replicated, this indicates that the P300 component can be used as an indirect marker of in-game performance, which might be of interest to game- or VR developers. Taken together, the data suggests that using the P300 as indirect marker of attentional allocation might be promising in future research but that the signal-to-noise ratio in this study might not have been high enough to find reliable effects. Although we used a within-subjects design to compare conditions, eighteen participants just might not have been enough.

When looking at brain oscillations, we found that there was a significant increase in high frequency activity (high beta and gamma; 21–100 Hz) for the VR conditions. There was no interaction with condition or an interaction between condition and device. We can think of two reasons for this finding. A first one relates to motor-related

brain activity: since it is known that beta and especially high-beta activity is associated with brain regions responsible for muscle activity [40], it is clear why we would observe more beta in the VR condition, where participants tended to look around more than during PC gaming. A second reason, however, directly relates to attentional focusing. Previous research has shown that gamma activity can reflect higher-order cognitive processes like mental effort and concentration [41]. Therefore, it is likely that participants were able to focus more on the game in VR because of the increased immersion and presence. An interesting discrepancy with previous research relates to the alpha frequency band (8–13 Hz). Whereas Núñez Castellar et al. [21] showed a significant alpha power increase in the flow condition, likely related to reward-related processes, we did not observe this difference. Another caveat we have to make is that we only looked at brain oscillations in posterior regions (Cz, Pz and Poz). Previous research has shown that different frequency bands reflect different processes depending on the location of the measurement, but we did not take this into account [e.g., 42]. Nevertheless, in contrast to measuring the P300 component, for which a secondary oddball task is required, looking at brain oscillations is a very straightforward and easy procedure. Therefore, we definitely think brain oscillations are a promising way to assess user experience during VR gaming in the future (given adapted HMDs that include EEG electrodes). Furthermore, there are still a lot of interesting research questions related to the aforementioned Synchronization theory of flow [12], which considers flow as a synchronization phenomenon of different attentional and reward networks in the brain. It would be interesting to investigate whether the observed increases in high beta and gamma are related to these attention- and reward-related networks. Finally, we measured the ECG signal and performed an analysis on the average heart rate per gaming session (beats-per-minute). This analysis did not reveal any interesting effects. However, there is some research suggesting a promising role for heart-rate-variability analyses in gaming research [43].

In sum, our research suggests that psychophysiological measures are promising tools to quantify attentional allocation in VR, but more research is needed to clarify whether and how this translates to flow.

Acknowledgements. We would like to thank Alexander Sels and Roel Mangelschots for their assistance with programming and modifying the game.

References

1. Rajesh Desai, P., Nikhil Desai, P., Deepak Ajmera, K., Mehta, K.: A review paper on oculus rift-a virtual reality headset. *Int. J. Eng. Trends Technol.* **13**, 175–179 (2014). <https://doi.org/10.14445/22315381/IJETT-V13P237>
2. Sanchez-Vives, M.V., Slater, M.: Opinion: from presence to consciousness through virtual reality. *Nat. Rev. Neurosci.* **6**, 332–339 (2005). <https://doi.org/10.1038/nrn1651>
3. Larson, M.J., Kaufman, D.A.S., Perlstein, W.M.: Neural time course of conflict adaptation effects on the Stroop task. *Neuropsychologia* **47**, 663–670 (2009). <https://doi.org/10.1016/j.neuropsychologia.2008.11.013>

4. Haluck, R.S., Krummel, T.M.: Computers and virtual reality for surgical education in the 21st century. *Arch. Surg.* **135**, 786–792 (2000)
5. Yen, C.-Y., Lin, K.-H., Hu, M.-H., Wu, R.-M., Lu, T.-W., Lin, C.-H.: Effects of virtual reality–augmented balance training on sensory organization and attentional demand for postural control in people with parkinson disease: a randomized controlled trial. *Phys. Ther.* **91**, 862–874 (2011). <https://doi.org/10.2522/ptj.20100050>
6. Gebara, C.M., de Barros-Neto, T.P., Gertsenchtein, L., Lotufo-Neto, F.: Virtual reality exposure using three-dimensional images for the treatment of social phobia. *Rev. Bras. Psiquiatr.* **38**, 24–29 (2016). <https://doi.org/10.1590/1516-4446-2014-1560>
7. Ermi, L., Mayra, F.: Challenges for pervasive mobile game design: examining players' emotional responses. In: *International Conference Proceedings Series*, vol. 265, pp. 371–372 (2005). <http://doi.acm.org/10.1145/1178477.1178554>
8. De Grove, F., Van Looy, J., Courtois, C.: Towards a serious game experience model: validation, extension and adaptation of the GEQ for use in an educational context. *Play. Play. Exp.* **10**, 47–61 (2010)
9. Kiili, K., De Freitas, S., Arnab, S., Lainema, T.: The design principles for flow experience in educational games. *Procedia Comput. Sci.* **15**, 78–91 (2012)
10. Csikszentmihalyi, M.: The flow experience and its significance for human psychology (1988)
11. Howe, M.J.A.: Flow - the psychology of happiness - Csikszentmihalyi. *Br. J. Educ. Psychol.* **63**, 539 (1993)
12. Weber, R., Tamborini, R., Westcott-Baker, A., Kantor, B.: Theorizing flow and media enjoyment as cognitive synchronization of attentional and reward networks. *Commun. Theory* **19**, 397–422 (2009). <https://doi.org/10.1111/j.1468-2885.2009.01352.x>
13. Chiang, Y.T., Lin, S.S.J., Cheng, C.Y., Liu, E.Z.F.: Exploring online game players' flow experiences and positive affect. *Turk. Online J. Educ. Technol.* **10**, 106–114 (2011)
14. Hoffman, D.L., Novak, T.P.: Flow online: lessons learned and future prospects. *J. Interact. Mark.* **23**, 23–34 (2009). <https://doi.org/10.1016/j.intmar.2008.10.003>
15. Sweetser, P., Wyeth, P.: GameFlow: a model for evaluating player enjoyment in games. *Comput. Entertain.* **3**, 3 (2005). <https://doi.org/10.1145/1077246.1077253>
16. Federoff, M.A.: Heuristics and usability guidelines for the creation and evaluation of fun in video games. FUN Video Games Thesis University Graduate School of Indiana University, 52 (2002). <http://doi.org/10.1.1.89.8294>
17. Davis, J.P., Steury, K., Pagulayan, R.: A survey method for assessing perceptions of a game: the consumer playtest in game design. *Game Stud* **5** (2005)
18. Desurvire, H., Caplan, M., Toth, J.A.: Using heuristics to evaluate the playability of games. In: *Extended Abstracts of the 2004 Conference on Human Factors and Computing Systems - CHI 2004*, p. 1509 (2004)
19. Nakamura, J., Csikszentmihalyi, M.: The Concept of Flow. *Flow and the Foundations of Positive Psychology*, pp. 239–263. Springer, Dordrecht (2014). https://doi.org/10.1007/978-94-017-9088-8_16. The Collected Works of Mihaly Csikszentmihalyi
20. Weber, R., Huskey, R.: Flow theory and media exposure: advances in experimental manipulation and measurement. Paper Accepted to the Annual Conference of the International Communication Association, London (2013)
21. Nuñez Castellar, E.P., Antons, J.-N., Marinazzo, D., van Looy, J.: Being in the zone: using behavioral and EEG recordings for the indirect assessment of flow. *PeerJ Prepr* **4**, e2482v1 (2016). <https://doi.org/10.7287/peerj.preprints.2482v1>
22. Luck, S.: An introduction to the event related potential technique (2005)
23. Buschman, T.J., Kastner, S.: From behavior to neural dynamics: an integrated theory of attention. *Neuron* **88**, 127–144 (2015)

24. Laufs, H., Krakow, K., Sterzer, P., Eger, E., Beyerle, A., Salek-Haddadi, A., Kleinschmidt, A.: Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest. *Proc. Natl. Acad. Sci.* **100**, 11053–11058 (2003). <https://doi.org/10.1073/pnas.1831638100>
25. Weber, R., Alicea, B., Mathiak, K.: The dynamic of attentional networks in mediated interactive environments. A functional magnetic resonance imaging study
26. Klasen, M., Weber, R., Kircher, T.T.J., Mathiak, K.A., Mathiak, K.: Neural contributions to flow experience during video game playing. *Soc. Cogn. Affect. Neurosci.* **7**, 485–495 (2012). <https://doi.org/10.1093/scan/nsr021>
27. Ulrich, M., Keller, J., Hoenig, K., Waller, C., Grön, G.: Neural correlates of experimentally induced flow experiences. *Neuroimage* **86**, 194–202 (2014). <https://doi.org/10.1016/j.neuroimage.2013.08.019>
28. Stanisor, L., van der Togt, C., Pennartz, C.M.A., Roelfsema, P.R.: A unified selection signal for attention and reward in primary visual cortex. *Proc. Natl. Acad. Sci.* **110**, 9136–9141 (2013). <https://doi.org/10.1073/pnas.1300117110>
29. Roto, V., Obrist, M., Väänänen-Vainio-Mattila, K.: User experience evaluation methods in academic and industrial contexts. *User Exp. Eval.* (2009). <http://doi.org/10.1.1.150.1764>
30. Mihajlovic, V., Grundlehner, B., Vullers, R., Penders, J.: Wearable, wireless EEG solutions in daily life applications: what are we missing? *IEEE J. Biomed. Health Inform.* **19**, 6–21 (2015). <https://doi.org/10.1109/JBHI.2014.2328317>
31. Sherry, J.L.: Flow and media enjoyment. *Commun. Theory* **14**, 328–347 (2004). <https://doi.org/10.1111/j.1468-2885.2004.tb00318.x>
32. Debener, S., Makeig, S., Delorme, A., Engel, A.K.: What is novel in the novelty oddball paradigm? Functional significance of the novelty P3 event-related potential as revealed by independent component analysis. *Cogn. Brain. Res.* **22**, 309–321 (2005). <https://doi.org/10.1016/j.cogbrainres.2004.09.006>
33. Fabiani, M., Kazmerski, V.A., Cycowicz, Y.M., Friedman, D.: Naming norms for brief environmental sounds: effects of age and dementia. *Psychophysiology* **33**, 462–475 (1996). <https://doi.org/10.1111/j.1469-8986.1996.tb01072.x>
34. Sherry, J.L., Rosaen, S., Bowman, N., Huh, S.: Cognitive skill predicts video game ability. In: Annual Meeting of the International Communication Association, Dresden (2006)
35. Delorme, A., Makeig, S.: EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004). <https://doi.org/10.1016/j.jneumeth.2003.10.009>
36. Lopez-Calderon, J., Luck, S.J.: ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front. Hum. Neurosci.* **8**, 213 (2014). <https://doi.org/10.3389/fnhum.2014.00213>
37. Polich, J.: P3A and P3B: towards an integrative theory. *Int. J. Psychophysiol.* **61**, 295 (2006)
38. Polich, J.: Updating P300: An integrative theory of P3a and P3b. *Clin. Neurophysiol.* **118**, 2128–2148 (2007)
39. Bombeke, K., Langford, Z.D., Notebaert, W., Nico Boehler, C.: The role of temporal predictability for early attentional adjustments after conflict. *PLoS One* **12**, e0175694 (2017). <https://doi.org/10.1371/journal.pone.0175694>
40. Muthukumaraswamy, S.D.: High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Front. Hum. Neurosci.* **7**, 138 (2013). <https://doi.org/10.3389/fnhum.2013.00138>
41. Pulvermüller, F., Birbaumer, N., Lutzenberger, W., Mohr, B.: High-frequency brain activity: Its possible role in attention, perception and language processing. *Prog. Neurobiol.* **52**, 427–445 (1997)

42. Kahana, M.J.: The Cognitive Correlates of Human Brain Oscillations. *J. Neurosci.* **26**, 1669–1672 (2006). <https://doi.org/10.1523/jneurosci.3737-05c.2006>
43. Castellar, E.N., Oksanen, K., Van Looy, J.: Assessing game experience: heart rate variability, in-game behavior and self-report measures. In: 2014 6th International Workshop on Quality of Multimedia Experience, QoMEX 2014, pp. 292–296 (2014)



M.I.N.D. Brain Sensor Caps: Coupling Precise Brain Imaging to Virtual Reality Head-Mounted Displays

Gyoung Kim¹, Joonhyun Jeon², and Frank Biocca^{1,3}(✉)

¹ M.I.N.D. Lab, Syracuse University, Syracuse, USA
{gmkim, fbiocca}@syr.edu

² Konkuk University, Seoul, Republic of Korea
Naturaljeon@konkuk.ac.kr

³ Media, Interface, and Network Design Lab, S.I. Newhouse School of Public Communications, Syracuse University, Syracuse, NY, USA

Abstract. Today, Virtual Reality (VR) and Augmented Reality (AR) are the new communication tools readily available to consumers. Because of the increasing availability of AR and VR, communication and neuroscience researchers are showing increasing interest in the use of VR systems for studies in collaboration, communication, and basic neuroscience. Beyond relying on self-reported or behavioral measures, psychophysiological or functional neuroimaging measurements sensing brain waves (e.g. EEG) or brain hemodynamics (e.g. fNIRS) are powerful techniques for measuring brain activity while interacting with virtual reality stimuli or environments. However, using these measures with virtual reality systems can be difficult due to physical and technical constraints. Both Functional Near-Infrared Spectroscopy (fNIRS) and Electroencephalography (EEG) need multiple channels to measure brain activity, a combination of cables and probes must be attached to a head cap. However, this setup obstructs wearing head-mounted display (HMD) in a VR environment and the challenge varies with the design of the HMD. To overcome these limitations, we introduce the design and development of the M.I.N.D. brain measurement cap specifically adapted for research with virtual reality system. We discuss the design process as well as the advantages and limitations of the current iterative design of the cap. Generally, we anticipate that this measurement system will expand the potential of influence of cognitive neuroscience contribute on VR research by making it easier for researchers to use a breadth of tools.

Keywords: Brain waves · EEG · fNIRS · Virtual Reality
Brain measurement method · Caps

1 Introduction

Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) are regarded as the new frontier in digital communication media and use of these new technology tools is rapidly increasing [1].

The development and increased availability of advanced brain measurement methods allows researchers to explore human behaviors and minds in different ways than were previously available. Researchers do not have to rely on entirely on the results from self-reported measurements, which have known limitations [2–4] or behavioral measures only. The emergence of neurophysiological measurements has opened up new fields and allowed others to grow because the tools allow us to explore previously unanswerable questions. For the reasons above, neurophysiological research has become increasingly popular.

This has been possible because the brain can be directly observed by non-invasive electrophysiological recording and the data shows accurate and high-resolution data.

2 A Review of Functional Neuroimaging and Electrophysiological Measurement

A cognitive neuroscience-based measurement is an advanced and improved technique to expand our knowledge of human behaviors and minds. These measurements are designed for non-invasively interpreting brain functioning at high temporal and/or spatial resolution. By capturing electric fields on the surface of the head induced by neuronal activity in the brain, we can specify human brain function.

This concept has been developed with the trust that the brain mainly controls physical and psychological activities [5]. For instance, this basic concept has widely contributed to develop brain-computer interfaces (BCI) using the brain as an addition input device for broader users including disable patients [6].

In addition, psychophysiological approaches to the brain have contributed to explore human minds deeper and wider without getting direct responses from participants. For example, neuroimaging studies have proven that increased amygdala activity when viewing fear [7], sad, angry [8] or happy facial expressions [7]. In addition, sensors are also able to measure mental workload [9], which is more effective and accurate than estimating workload from post-test questionnaires. The reason is first, it does not instantly reflect the result at the time of exposure (delay in response). In most cases, the questionnaire is completed after exposure, not during exposure of stimuli. Therefore, to complete a post-test questionnaire, a participant needs to recall the time of the experiment. Secondly, the result from the 5 or 7 scale questionnaires does not represent the details of participants' minds and thoughts because human's psychological process is not readily available for introspection. On the contrary, the brain imaging tool using a psychophysiological method continuously captures user's cognitive process on a millisecond scale. For these reasons, psychophysiological measurements have been widely used even though it requires more time and cost to set the experimental environment up.

Then, what types of brain waves we can measure and what do they represent? There are technically different methods to measure brain waves, and each method has a different approach to explore human minds.

There are several physiological measurements (i.e. PET, fMRI or TMS), however we discuss only methods that measure brain waves with a head cap and sensors since a

purpose of this paper is to suggest a new design of the cap for brain wave measurement system.

2.1 Electroencephalogram (EEG)

In contrast to other physiological measures such as heart rate, skin-conductance level, or facial electromyography (EMG), Electroencephalogram (EEG) is a direct measure of central nervous system activity from the active brain. In particular, the EEG can see the changes in electrical discharge of cortical neuronal populations; it measures the capacity or performance of cortical information [10]. In other words it can measure the degree of brain concentration [11]. It can measure the arousal dimension of human emotions from peripheral signals [12] which may include alpha, theta and/or frequencies greater than 16 Hz [13]. In addition, it is ease of use and low set-up cost. It generally uses from 2–256 channels to see each area of the brain by wearing a head cap and putting measuring sensors (electrodes) into grommets over the cap. Grimes et al. showed 99% accuracy in working memory states (WM) and four WM states with up to 88% accuracy with EEG [14]. Because of this accuracy, military also uses EEG to monitor pilots' mental states while they are in the air [15] (Fig. 1).



Fig. 1. EEG head cap (Source: Biopac Systems)

2.2 Functional Near-Infrared Spectroscopy (fNIRS)

Functional near-infrared spectroscopy (fNIRS) is another advanced neuroimaging technology for mapping the functioning human cortex. It was originally designed for medical use and has been first clinically used in the mid-1980s. However, it is now playing an important role in both neuroscience and communication research [16]. The fNIRS uses a measurement of concentration changes in both oxygenated and deoxygenated hemoglobin (Hb). It is also a non-invasive imaging method to see the changes in blood oxygenation, that can represent levels of brain activation while EEG captures electrical waves associated with the activation potential of neurons. The fNIRS uses optical fibers placed on the scalp that send light in the wavelength range of 650–850 nm in to the targeted area on the head, where the infrared light is re-emitted based on the amount of oxygenated and deoxygenated hemoglobin from the tissue during brain function from multiple channels (up to 256).

This method has been proven by many researchers as a valid method to measure hemodynamic levels originating from prefrontal cortex (PFC) activation [17] in emotion induction [18] and cognitive functions such as problem solving or memory related to mental workload [19] (Fig. 2).



Fig. 2. fNIRS head cap (Source: Hitachi Medical Systems)

3 Benefits of Psychophysiological Measures to Understand New Media

The brain imaging tools we discussed above have several common advantages for research. First, they accurately measure human mental states or workload with sub-second or even millisecond level precision. Secondly, those are popular, non-invasive and widely proven techniques in research. Third, EEG or fNIRS can measure brain waves in more ecologically valid conditions than other sensors. For example, using a PET or fMRI require participants to be stationary or lie in restricted positions.

For those reasons, EEG and fNIRS have been widely used in human computer interaction (HCI) studies, including usability testing and user-centered design for better human performance [9]. In addition, EEG and fNIRS are also suitable for new media research when scientists are interested in the cognitive mechanisms behind media effects [20, 21].

3.1 Virtual Reality

The concept of Virtual Reality (VR) is to experience a certain unreal environment as real. With the technological development in graphics, it is possible to depict a realistic environment in 3D. By wearing Head-mounted display (HMD), it is possible for a user to be in a virtual world without seeing other real objects near the user that may distract experiencing immersive VR.

In addition, other functionalities that increase users' emersion, such as head-tracking or motion-tracking system, make a virtual environment more realistic so that users feel "they are actually there" [22–24]. Therefore, in recent years

communication researchers and sociologists have studied to see their minds when they interact with other objects or people in VR [25, 26].

3.2 Challenges in Recording EEG and fNIRS in Virtual Reality

As we discussed above, it is essential to see users' cognition when they interact in VR. It is essential to see how users accept and process information they get and interact with in virtual environments.

However, combining VR system and EEG or fNIRS is difficult because the HMD mounts interfere with the probes of the devices. There were some studies that see some social effects of virtual reality with electroencephalography (EEG) or fNIRS previously [27, 28], however they just see the effect of virtual object shown in a 2D display, not an HMD.

As shown in the Fig. 3, both HTC VIVE and Oculus Rift use a 3-axis headband for a perfect fit. In addition, the headband should be tightened enough to cut off the light from outside of the HMD, so the user can fully focus in virtual environment. However, once the user puts the brain measurement cap on, they are unable to wear the HMD over the cap since the cables and sensors attached to the cap block placement of the HMD. Since the HMD head bands are made with a flexible rubber, it puts pressure on the sensors. This pressure displaces the probes, which results in poor sensor contact with the head and consequently, poor data quality. Also, to measure prefrontal cortex activity in VR, the HMD cannot be worn correctly because the HMD obstructs the area over the prefrontal cortex. If the HMD is worn below the cap, then the HMD lens and eye are not aligned and centered, that finally results in a failure of setting up the virtual environment. Figure 4 shows the experimental setup of fNIRS. As shown in the picture, it is impossible to wear the HMD over the cap because of cables.



Fig. 3. Virtual reality head-mounted display: Oculus Rift and HTC Vive (Source: Oculus and HTC)

For those restrictions, it has always been a difficulty for scholars to research an immersive virtual environment using sensors like the fNIRS.

3.3 Previous Solutions Suggested by Other Researchers

Figure 5 shows a method to use fNIRS in VR setup suggested by Seraglia et al. [29] in 2011. They developed a VR cap that does not interfere fNIRS cables. However, the configuration limits the regions of the brain that can be measured, which functionally prohibits research on topics like cognitive load. For instance, prefrontal cortex activity



Fig. 4. Example of brain measurement setup

seeing many higher cognitive functions [30] cannot be measured with this setup. Many researchers investigated prefrontal cortex activity in the simulated environment created with computer graphics [31, 32]. Therefore, it was highly needed to develop a new brain measurement cap to solve this limitation.

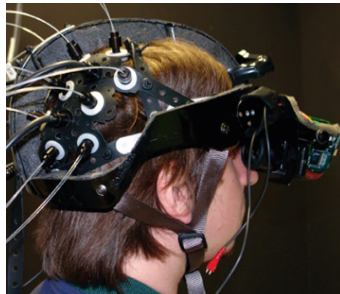


Fig. 5. Suggested solution to use fNIRS in VR by Seragila et al. [29]

4 M.I.N.D. Brain Cap, Measurement for VR Research: Design and Iterations

4.1 M.I.N.D. Brain Cap, Initial Design

To solve issues and restrictions of brain measurements in VR discussed above, we designed an innovative head cap that makes it possible to wear the head-mounted display over the brain measurement cap (Figs. 6 and 7).

There are several advantages to this design. First, researchers do not need to limit the number of channels to use because the cap is designed to make most channels available in the VR experimental setup.

Secondly, this cap saves time in setting up an experimental environment because subjects need to wear only one cap that incorporate (1) the HMD, and (2) the brain sensor array. In a previous setup, a subject wears the brain measurement cap first and

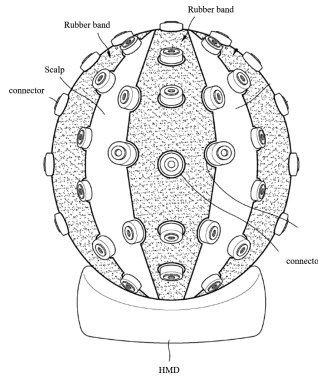


Fig. 6. New design of the brain measurement cap for immersive virtual reality system (patent pending)

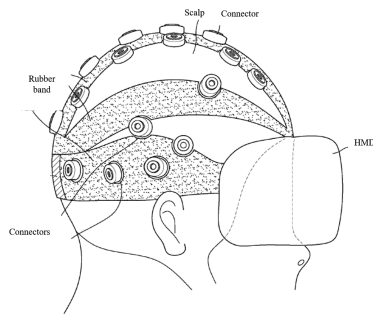


Fig. 7. Side view of the new brain measurement cap

then wear the HMD. Third, the cap design works with both EEG and fNIRS. Sensor sockets can be easily replaced for each experimental measurement (Fig. 8).



Fig. 8. Back side view of the M.I.N.D. brain measurement cap

However, there are also limitations of the cap. First, even though the main material of the cap is made of rubber, expansion of the cap could not cover all head sizes. For instance, optical signals showed poor in the front forehead area when the cap was worn on a smaller head. Secondly, connectors to attach HMDs are not universal, therefore we need to design a cap for each HMD (e.g. Oculus rift version, or HTC VIVE version). Thirdly, modifying the channel configuration and probe placement is difficult.

4.2 M.I.N.D. Brain Cap: Second Version

To solve the issues of the first version of the cap, we made several design modifications. First, we choose a modular design to address the size issue of the cap.

As shown in Fig. 9, the cap consists of three different parts and each part can be easily connected to other modules (probe holders) with several plastic clips. By doing so, a researcher can easily find a perfect fit (size) for participant's head by configuring different size of cap modules. Therefore, we have a higher probability of accurate probe placement and better-quality data. Secondly, the researcher can easily rearrange the measurement area by choosing various size of modules, which is a standard option for non-VR research. For instance, standard configuration of the modular consist of 4×4 and 3×3 channels. If a researcher wants to concentrate on a certain area, then he or she can use 4×4 , and choose 3×3 for the rest area in fNIRS (See Fig. 10 for details).

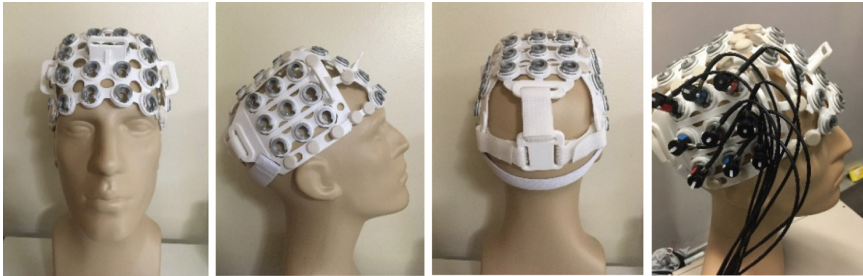


Fig. 9. Second version of the M.I.N.D. brain measurement cap (modular design)



Fig. 10. Various sizes of probe holders (source: Hitachi Medical Systems)

5 Conclusion

The emergence of VR and AR as a readily available communication tool and media interface has pushed researchers to begin exploring the cognitive neuroscience associated with VR/AR experiences. However, conducting studies is difficult because of technical challenges associated with integrated HMDs and sensors. The M.I.N.D. brain measurement cap is the first attempt at addressing these problems; we hope our design makes it easier for researchers to combine VR/AR and EEGs and fNIRS.

References

1. Ryan, M.-L.: *Narrative as Virtual Reality 2: Revisiting Immersion and Interactivity in Literature and Electronic Media*. Johns Hopkins University Press, Baltimore (2015)
2. Donaldson, S.I., Grant-Vallone, E.J.: Understanding self-report bias in organizational behavior research. *J. Bus. Psychol.* **17**, 245–260 (2002). <https://doi.org/10.1023/A:1019637632584>
3. McDonald, J.D.: Measuring personality constructs: the advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire* **1**, 1–19 (2008)
4. Stone, A.A., Bachrach, C.A., Jobe, J.B., Kurtzman, H.S., Cain, V.S.: *The Science of Self-report: Implications for Research and Practice*. Lawrence Erlbaum, Mahwah (1999)
5. Coles, M.G.H.: Modern mind-brain reading: psychophysiology, physiology, and cognition. *Psychophysiology* **26**, 251–269 (1989). <https://doi.org/10.1111/j.1469-8986.1989.tb01916.x>
6. Girouard, A., Solovey, E.T., Hirshfield, L.M., Peck, E.M., Chauncey, K., Sassaroli, A., Fantini, S., Jacob, R.J.K.: From brain signals to adaptive interfaces: using fNIRS in HCI. In: *Brain-Computer Interfaces Applying our Minds to Human-Computer Interaction*, pp. 221–237 (2010)
7. Breiter, H.C., Etcoff, N.L., Whalen, P.J., Kennedy, W.A., Rauch, S.L., Buckner, R.L., Strauss, M.M., Hyman, S.E., Rosen, B.R.: Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* **17**, 875–887 (1996). [https://doi.org/10.1016/s0896-6273\(00\)80219-6](https://doi.org/10.1016/s0896-6273(00)80219-6)
8. Johnson, B.T., Eagly, A.H.: Effects of involvement on persuasion: a meta-analysis. *Psychol. Bull.* **106**, 290–314 (1989). <https://doi.org/10.1037/0033-2909.106.2.290>
9. Hirshfield, L.M., Chauncey, K., Gulotta, R., Girouard, A., Solovey, E.T., Jacob, R.J.K., Sassaroli, A., Fantini, S.: Combining electroencephalograph and functional near infrared spectroscopy to explore users' mental workload. In: Schmorrow, D.D., Estabrooke, I.V., Grootjen, M. (eds.) *FAC 2009. LNCS (LNAI)*, vol. 5638, pp. 239–247. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02812-0_28
10. Klimesch, W.: EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* **29**, 169–195 (1999). [https://doi.org/10.1016/S0165-0173\(98\)00056-3](https://doi.org/10.1016/S0165-0173(98)00056-3)
11. Liu, N.H., Chiang, C.Y., Chu, H.C.: Recognizing the degree of human attention using EEG signals from mobile sensors. *Sens. (Switz.)* **13**, 10273–10286 (2013). <https://doi.org/10.3390/s130810273>
12. Chanel, G., Kronegg, J., Grandjean, D., Pun, T.: Emotion assessment: arousal evaluation using EEG's and peripheral physiological signals. In: Günsel, B., Jain, A.K., Tekalp, A.M., Sankur, B. (eds.) *MRCSS 2006. LNCS*, vol. 4105, pp. 530–537. Springer, Heidelberg (2006). https://doi.org/10.1007/11848035_70

13. Bonnet, M.H., Carley, D., Consultant, M.C., Consultant, P.E., Chairman, C.G., Harper, R., Hayes, B., Hirshkowitz, M., Keenan, S., Consultant, M.P., Roehrs, T., Smith, J., Weber, S., Westbrook, P., Administrative, A., Bruce, S.: EEG arousals: scoring rules and examples: a preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorders Association. *Sleep* **15**, 173–184 (1992)
14. Grimes, D., Tan, D.S., Hudson, S.E., Shenoy, P., Rao, R.P.N.: Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In: Proceedings of the 26th SIGCHI Conference on Human Factors in Computing Systems, p. 835 (2008). <https://doi.org/10.1145/1357054.1357187>
15. Karim, H., Schmidt, B., Dart, D., Beluk, N., Huppert, T.: Functional near-infrared spectroscopy (fNIRS) of brain function during active balancing using a video game system. *Gait Posture*. **35**, 367–372 (2012). <https://doi.org/10.1016/j.gaitpost.2011.10.007>
16. Liu, N., Mok, C., Witt, E.E., Pradhan, A.H., Chen, J.E., Reiss, A.L.: NIRS-based hyperscanning reveals inter-brain neural synchronization during cooperative Jenga game with face-to-face communication. *Front. Hum. Neurosci.* **10** (2016). <https://doi.org/10.3389/fnhum.2016.00082>
17. Sato, H., Yahata, N., Funane, T., Takizawa, R., Katura, T., Atsumori, H., Nishimura, Y., Kinoshita, A., Kiguchi, M., Koizumi, H., Fukuda, M., Kasai, K.: A NIRS–fMRI investigation of prefrontal cortex activity during a working memory task. *Neuroimage* **83**, 158–173 (2013). <https://doi.org/10.1016/j.neuroimage.2013.06.043>
18. Herrmann, M.J., Ehlis, A.C., Fallgatter, A.J.: Prefrontal activation through task requirements of emotional induction measured with NIRS. *Biol. Psychol.* **64**, 255–263 (2003). [https://doi.org/10.1016/S0301-0511\(03\)00095-4](https://doi.org/10.1016/S0301-0511(03)00095-4)
19. Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., Chance, B.: Functional optical brain imaging using near-infrared during cognitive tasks. *Int. J. Hum. Comput. Interact.* **17**, 211–227 (2004)
20. Bolls, P.D., Lang, A., Potter, R.F.: The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements. *Commun. Res.* **28**, 627–651 (2001). <https://doi.org/10.1177/009365001028005003>
21. Park, B.: Psychophysiology as a tool for HCI research: promises and pitfalls. In: Jacko, J.A. (ed.) *HCI 2009*. LNCS, vol. 5610, pp. 141–148. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02574-7_16
22. Biocca, F.: The Cyborg’s dilemma: progressive embodiment in virtual environments minding the body, the primordial communication medium. *JCMC* **3**, 1–29 (1997). <https://doi.org/10.1111/j.1083-6101.1997.tb00070.x>
23. Heeter, C.: Being there: the subjective experience of presence. *Presence* **1**, 262–271 (1992). <https://doi.org/10.1109/VRAIS.1995.512482>
24. Reeves, B., Nass, C.: *How People Treat Computers, Television, and New Media Like Real People and Places* (1998)
25. McCabe, K., Houser, D., Ryan, L., Smith, V., Trouard, T.: A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl. Acad. Sci.* **98**, 11832–11835 (2001). <https://doi.org/10.1073/pnas.211415698>
26. Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., Kilts, C.D.: A neural basis for social cooperation. *Neuron* **35**, 395–405 (2002). [https://doi.org/10.1016/S0896-6273\(02\)00755-9](https://doi.org/10.1016/S0896-6273(02)00755-9)
27. Baumgartner, T., Valko, L., Esslen, M., Jäncke, L.: Neural correlate of spatial presence in an arousing and noninteractive virtual reality: an EEG and psychophysiology study. *CyberPsychol. Behav.* **9**, 30–45 (2006). <https://doi.org/10.1089/cpb.2006.9.30>

28. Schilbach, L., Koubeissi, M.Z., David, N., Vogeley, K., Ritzl, E.K.: Being with virtual others: Studying social cognition in temporal lobe epilepsy. *Epilepsy Behav.* **11**, 316–323 (2007). <https://doi.org/10.1016/j.yebeh.2007.06.006>
29. Seraglia, B., Gamberini, L., Priftis, K., Scatturin, P., Martinelli, M., Cutini, S.: An exploratory fNIRS study with immersive virtual reality: a new method for technical implementation. *Front. Hum. Neurosci.* **5** (2011). <https://doi.org/10.3389/fnhum.2011.00176>
30. Miller, E.K., Cohen, J.D.: An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001). <https://doi.org/10.1146/annurev.neuro.24.1.167>
31. Calhoun, V.D., Pekar, J.J., McGinty, V.B., Adali, T., Watson, T.D., Pearlson, G.D.: Different activation dynamics in multiple neural systems during simulated driving. *Hum. Brain Mapp.* **16**, 158–167 (2002). <https://doi.org/10.1002/hbm.10032>
32. Matsuda, G., Hiraki, K.: Sustained decrease in oxygenated hemoglobin during video games in the dorsal prefrontal cortex: a NIRS study of children. *Neuroimage* **29**, 706–711 (2006). <https://doi.org/10.1016/j.neuroimage.2005.08.019>



Assessing Operator Psychological States and Performance in UAS Operations

Jinchao Lin^(✉), Gerald Matthews, Lauren Reinerman-Jones,
and Ryan Wohleber

Institute for Simulation and Training (IST), University of Central Florida (UCF),
Orlando, FL, USA

{jlin, gmatthew, lreiner, rwohlebe}@ist.ucf.edu

Abstract. Assessment for understanding, predicting, and improving human performance and system design is a key for human-computer interaction (HCI) research. Assessments can be behavioral, physiological, performance-based, and phenomenological. Assessments are important in a variety of domains, including unmanned vehicle operations, human-robot teaming, nuclear power plant operations, etc. This paper will discuss assessment approaches in the domain of unmanned aerial systems (UAS) operations to identify and quantify explanatory constructs, such as psychological states, workload, and performance. It will also discuss implications for evaluating improvements in human performance in UAS operations. Specifically, this paper will examine metrics that can be utilized to gauge the impact of demand factors on workload, task performance, operator dependence on automation, and stress response.

Keywords: UAS · Workload · Stress · Performance · Assessment

1 Introduction

Nowadays, machines with advanced technology are capable of performing varieties of jobs which are currently performed by humans. Although some machines are designed to do their jobs automatically, human operators, at least in a supervisory role, are still required in many cases. The development of technology makes complex human-machine systems possible and brings numerous benefits, but it may also lead to human factors challenges in terms of operator performance. Assessing operators' psychological states and performance is essential for understanding, predicting, and improving human performance and system design in those complex human-machine systems. This paper will take a simulation study of unmanned aerial systems (UAS) operations as an example to discuss assessment approaches for identifying and quantifying explanatory constructs. It focuses especially on the challenges of measuring various facets of overload, such as psychological states, subjective workload, and performance impairment.

In current UAS mission configurations, operators are often more at risk from underload than overload, due to requirements to loiter for extended periods until a target appears. However, workload issues may change as future UAS will allow multi-aircraft control (MAC) by a single operator. This is anticipated to be a

particularly time-critical, and cognitively demanding, form of multi-tasking work [1, 2]. Automation of system functions is essential for MAC, but the necessity of keeping the operator in the loop and capable of coordinating multiple vehicles limits the extent to which automation can mitigate the increased workload associated with multi-tasking. Evaluating system design and operator competencies requires assessment methodologies that can detect various expressions of overload, including excessive subjective workload, stress, and performance impairment. One of the least understood metrics for overload is change in trust [3]. Appropriate trust in automation may be especially important in overload situations, but there is rather little evidence on the extent to which trust measures converge or diverge with those for operator overload.

1.1 Workload and Stress in UAS Operation

UAS operations often involve considerable task demand variation which may be accompanied by both stress, and changes in operator performance. In Hancock and Warm's [4] theoretical model for stress and performance, individuals can adapt effectively to some levels of task demands without showing significant performance decrement. However, both extreme overload and underload could result in failures in such adaptation. In terms of MAC, overload from sources including monitoring displays, flight control, navigation, communication, and mission management is the primary challenge [5]. As the number of vehicles controlled increases, operators may become more vulnerable to overload due to the limited resources taxed by multiple demand factors [6].

Automation plays a critical role in mitigating the workload associated with MAC by keeping cognitive demands to a manageable level [5]. Calhoun et al. [1] showed that automation benefits for performance transferred from the subtask automated to additional subtasks, implying that automation may free some general attentional resources. However, automation may fail to mitigate workload if it is poorly designed or used inappropriately [7].

The impact of automation on workload may also depend on the level of automation (LOA), i.e., the extent of the tradeoff between operator control and delegation of control to the machine [8]. Generally, higher LOAs reduce operator workload, but may impair vigilance and situation awareness. MAC studies typically envisage intermediate levels of control such that the operator must check and possibly over-ride the recommendations of the automation. Effects of LOA on performance and on trust in automation are mixed in empirical studies: the optimal LOA may depend on task demands and other aspects of system configuration [1, 9].

Stress commonly accompanies high workload [10]. Indeed, UAS operators report that task demand factors such as interface difficulties and inefficiencies in control procedures contribute to stress, along with occupational health factors [11]. However, stress may be less than in conventional flying. Skilled performers are often able to mitigate task-induced stress by developing strategies that prevent catastrophic performance failures and strategies for emotion-regulation [12].

In an empirical study using a MAC simulation, Wohleber et al. [13] examined the impact of task demand on both subjective and physiological stress indices. High task demand produced substantial increases in workload and distress, but only affected

certain stress indices. Specifically, demand increased high-frequency activity in the electroencephalogram (EEG), but did not affect cardiac response. These results suggested that high demand produced a relatively subtle, “cognitive” form of stress, rather than the classic fight-or-flight response, which would have elevated heart rate [13].

The present study focused on assessing multiple impacts of task demands during MAC operations. The simulation required multi-tasking with support from automation. It was configured to prioritize performance of surveillance activities common in UAS operations. Sustained monitoring often imposes high workload leading to depletion of processing resources and vigilance decrement [14], as well as subjective distress and loss of task engagement [15]. The study evaluated the extent to which the different indicators of overload converged under high task demands, as well as investigating the overload effect on dependence on automation.

1.2 Assessment of Operator Response to Cognitive Demands

Strain on UAS operators takes a variety of forms, influencing multiple objective and subjective responses. In one of the ISR tasks used in the present experiment, participants were asked to monitor displays to identify targets and take actions based on the discrimination of enemy and friendly tanks. In this case, accuracy, which is the percentage of correct actions, can be a straightforward assessment reflecting operator performance. However, the multi-component nature of the task also suggests that demand factors might influence subtask prioritization and possible neglect of subtasks [16]. Automated systems also raise the issue of choosing metrics for trust and reliance on the automation in decision-making. Trust is typically seen as a quality of operator state antecedent to behavior [3]. Behavioral indices seek to capture the different ways in which operator use of automation is suboptimal [17]. In the context of automated alarms, performance is typically assessed in terms of two measures [18]: compliance (taking action following the alarm) and reliance (taking no action in the absence of an alarm). However, this distinction may not be applicable to routine binary stimulus discrimination, where neither option constitutes an alarm requiring urgent response. Therefore, in this study, the metric of dependence on automation, the extent to which the operator follows the recommendations of the automation across all trials [19], was adopted as a behavior index for trust.

Workload and stress can be assessed using objective, physiological response and subjective scales [20]. The current study used the subjective measures sensitive to task demands in previous studies of simulated UAS [21] and unmanned ground vehicle (UGV) [22] operations: the NASA-TLX [23] workload measure, and the Dundee Stress State Questionnaire that assesses affective, motivational and cognitive aspects of subjective state (DSSQ) [10]. The DSSQ represents a multidimensional perspective on stress: task demands and environmental pressures elicit a range of qualitatively different subjective state responses [24].

The present research focused on the three higher-order DSSQ factors extracted from the primary dimensions: distress, task engagement, and worry. Distress is driven by a sense of being overloaded and lack of control over the task environment, coupled with negative affect [25]. Task demand manipulations often influence distress and workload concurrently [24, 25], and the distress scale correlates with NASA-TLX

workload [10]. In unmanned vehicle studies, multi-tasking demands elevate distress substantially [17, 22]. Task engagement reflects energy, task motivation, and alertness. Operators managing UAS may be prone to both upwards and downwards shifts in task engagement. For example, complex, challenging tasks including game-like scenarios involving unmanned vehicle control tend to elevate task engagement [2, 25]. In a UGV study, Abich et al. [22] found that both multi-tasking and higher event rates produced moderate elevations of engagement. On the other hand, prolonged, monotonous tasks such as vigilance typically lower task engagement substantially [24, 26]. Guznov et al. [27] found moderate-magnitude engagement decline during a 30-min UAS surveillance mission, consistent with reports of monotony and fatigue during real-life missions of substantially longer durations [28]. The third higher-order DSSQ dimension, worry, corresponds to self-focused attention, low self-esteem, and high cognitive interference. In unmanned vehicle studies, it tends to be less sensitive to task demand manipulations than distress and task engagement [22]. Poorly designed automation that threatens the operator's sense of personal competence might elevate worry [24]. Worry also overlaps with mind-wandering which may develop during prolonged operations [28].

1.3 Aims and Hypotheses

The study aimed to apply a multivariate assessment strategy [20] to profile the workload, stress and performance changes associated with high task demands during simulated UAS operation. We tested hypotheses derived from the Standard Capacity Model (SCM) [12]. The model captures the typical assumption of workload researchers that overload is associated with an insufficiency of general attentional resources to meet demands for processing. Lack of resources drives objective performance impairment and high subjective workload and distress. The effects of resource insufficiency on trust and automation-dependence have not been much researched; however, Parasuraman and Manzey [29] theorized that high task demands would tend to increase dependence on automation as a compensatory strategy (depending also on the configuration of the automation). In the limiting case, the different responses to overload would be interchangeable as indices of a general overload or resource insufficiency syndrome. However, we expected that we would in fact find divergences between different measures as in previous, related studies [21, 22, 30]. We assessed the strength of impact of task demands on dependent variables in terms of Cohen's effect size measure d . Conventionally, d values of .2, .5, and .8 correspond to small, medium and large effect sizes. The study also manipulated LOA. In the present study, LOA was of interest primarily as a potential moderator of task demand effects, but we report its effects briefly.

2 Method

2.1 Participants and Experimental Design

A total of 101 college students (42 men, 59 women, $M_{\text{age}} = 18.95$, $SD = 1.80$) participated this study for course credit. Participants were healthy individuals between 18

and 40 years old representing the age group and educational level of the enlisted military service core that may be recruited for future UAS operations. Participants who may be vulnerable to adverse reactions, such as excessive stress, resulting from the test environment were excluded. All participants reported having normal or corrected to normal vision, color vision, normal hearing, and English fluency.

A 2 (task demand: high versus low) \times 2 (LOA: management-by-consent versus management-by-exception) between-subjects factorial design was adopted in this study to assess demand factor impacts and operator performance. Twenty-six participants took part in the low task demand/management-by-exception condition; there were twenty-five participants in the other three conditions.

2.2 Simulation

The ALOA (Adaptive Levels of Autonomy) multi-UAS research test bed developed by OR Concepts Applied [1] was used for the study. This simulation supports task manipulations representing UAS operations in needed complexity and realism (see Fig. 1). Nine tasks, including target allocation and rerouting, two surveillance tasks, aircraft identification, status monitoring, decision making, and information retrieval, were designed to represent the task demands for a single operator managing a fleet of four aircrafts with an automation aid at the same time. The LOAs were varied in two intermediate levels with high reliability (correct 80% of the time) to support the two primary surveillance tasks, Image Analysis and Weapon Release authorization. Management-by-consent required participants to accept or change the option recommended by the automation. Alternatively, with management-by-exception, the system was set to act on the option recommended by the automation automatically unless a different option was selected before the availability of operator response was timed out (30 or 20 s based on tasks).

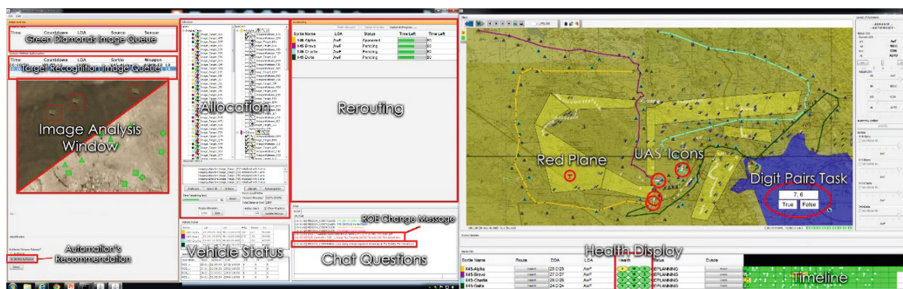


Fig. 1. Task interface for MAC operation in the ALOA UAS simulation

Task frequencies of the secondary tasks were manipulated to create task demand variation across conditions in one hour long experimental scenarios whereas task frequencies of two primary surveillance tasks were held constant. There were 6 tasks or 14 tasks per minute to induce low and high task demand respectively. Most secondary tasks required responses to visual or audio signals, searching and retrieving

information, or comparing digit pairs. Image Analysis and Weapon Release authorization tasks were timed for 30 and 20 s respectively. The task response would be recorded as a “miss” if there is no operator response before task availability was timed out.

In the Image Analysis task, images with an overlay of 19–26 green symbols varying in shapes, including diamonds, squares, circles, and triangles, were shown in the task panel. Participants were asked to count the number of diamonds and select the number from eight options. The automated aid system recommended one from the eight options by highlighting it. The reliability of the automation was set to be 80% correct.

In the Weapon Release authorization task, participants were asked to distinguish hostile tanks from allied tanks and detect whether the hostile tanks in given picture were correctly marked. The tanks differed in body width and barrel length subtly. The pictures were degraded in quality to increase the difficulty in discrimination. The automation aid system recommended one option from “authorize” or “do not authorize”. Also, reliability was set to be 80% correct.

2.3 Subjective Measures

Demographics Questionnaire. We developed a 21-item demographics questionnaire with questions on a range of biographical information, including age, gender, health status, education level, computer expertise, and gaming experience and expertise.

Stress state. The short, 21-item version of the Dundee Stress State Questionnaire (DSSQ) [25] was used to measure three higher order dimensions of subjective states in terms of task engagement, distress, and worry. Participants rated the accuracy of statements about their subjective states using a 5-point Likert.

Workload. The NASA-Task Load Index [23] consists of six 0–100 rating scales, including mental demand, physical demand, temporal demand, performance, effort, and frustration. Overall workload was calculated as the unweighted mean of the six ratings.

2.4 Performance-Based Measures

Three performance metrics for the two high priority surveillance tasks, Image Analysis and Weapon Release authorization, were analyzed. Accuracy was defined as the percentage of correct responses. Dependence on automation was defined as the percentage of trials on which the participant followed the recommendation from the automation. Neglect was defined as the frequency of items that appeared in the task window but were not opened by the participant. Detailed performance metric formulas for Image Analysis and Weapon Release authorization tasks are listed in Tables 1 and 2. Correct responses were labeled as “hit” (agree with automation) and “correct rejection” (disagree with automation). Incorrect answers were labeled as “near/far miss” (agree with automation) and “false alarm” (disagree with automation). In low LOA condition, if no action was taken before time-out, the task was recorded as a “true miss”.

Table 1. Performance metrics in the Image Analysis task

	Formula
Low LOA	
Accuracy	$\frac{Hit + CorrectRejection}{Hit + CorrectRejection + NearMiss + FarMiss + FalseAlarm + TrueMiss} \times 100\%$
Dependence on Automation	$\frac{Hit + NearMiss + FarMiss}{Hit + CorrectRejection + NearMiss + FarMiss + FalseAlarm + TrueMiss} \times 100\%$
Neglect	Number of tasks which the participant never opened
High LOA	
Accuracy	$\frac{Hit + CorrectRejection}{Hit + CorrectRejection + NearMiss + FarMiss + FalseAlarm} \times 100\%$
Dependence on Automation	$\frac{Hit + NearMiss}{Hit + CorrectRejection + NearMiss + FarMiss + FalseAlarm} \times 100\%$
Neglect	Number of tasks which the participant never opened

Table 2. Performance metrics in the Weapon Release authorization task

	Formula
Low LOA	
Accuracy	$\frac{Hit + CorrectRejection}{Hit + CorrectRejection + NearMiss + FalseAlarm + TrueMiss} \times 100\%$
Dependence on Automation	$\frac{Hit + NearMiss}{Hit + CorrectRejection + NearMiss + FalseAlarm + TrueMiss} \times 100\%$
Neglect	Number of tasks which the participant never opened
High LOA	
Accuracy	$\frac{Hit + CorrectRejection}{Hit + CorrectRejection + NearMiss + FalseAlarm} \times 100\%$
Dependence on Automation	$\frac{Hit + NearMiss}{Hit + CorrectRejection + NearMiss + FalseAlarm} \times 100\%$
Neglect	Number of tasks which the participant never opened

2.5 Procedure

Following an informed consent procedure, participants were instructed to complete the pre-task survey set, including the Demographic Questionnaire, and the pre-task DSSQ. After completing pre-task surveys, training started with an introduction using Power-Point slides, followed by a live simulation demonstration and hands-on practice. A “cheat sheet” about all the tasks was provided for quick reference. Training took approximately 60 min. Participants have to be qualified in the practice. Before the experimental task, researcher repeated instructions for simulation controls briefly and emphasized task priorities. The task ran for 60 min. Finally, participants were instructed to complete the post-task DSSQ and NASA-TLX, prior to debriefing. All the sessions in total were completed within three hours.

3 Results

3.1 Workload

Bonferroni-corrected *t*-tests were run to test the effects of experimental manipulations. It was confirmed that workload (NASA-TLX global workload) was significantly higher in high task demand conditions ($M = 57.1$) than in low task demand conditions ($M = 46.2$), $t(99) = -3.52$, $p = .001$, $d = .70$. According to NASA-TLX, the manipulation of task demand successfully elicited higher workload in all aspects, including mental demand, $t(99) = -1.78$, $p = .079$, $d = .35$; physical demand, $t(99) = -3.77$, $p < .01$, $d = .75$; temporal demand, $t(99) = -2.43$, $p < .05$, $d = .48$; effort, $t(99) = -2.47$, $p < .05$, $d = .49$, and frustration, $t(99) = -2.73$, $p < .01$, $d = .54$, in high task demand conditions (see Fig. 2). However, there was no difference in self-reported performance, $t(99) = -.21$, $p = .835$.

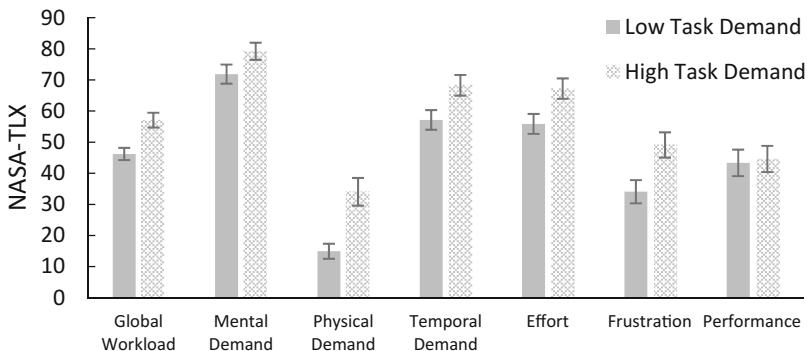


Fig. 2. NASA-TLX workload factor ratings in low/high task demand conditions. Error bars represent standard errors.

Bonferroni-corrected *t*-tests were also computed to check the impact of LOA manipulations. No significant self-rated workload differences were found between different LOA conditions.

3.2 Stress State

A series of $2 \times 2 \times 2$ (LOA \times task demand \times pre- vs. post-task) mixed-model ANOVAs were run for each stress state factors, including task engagement, distress, and worry, to test the effects of experimental manipulations on subjective states. The effect sizes (Cohen's d) for the stress state changes in different task demand conditions are shown in Fig. 3. The results from ANOVA for task engagement showed a near significant interaction between pre-/post-task and task demand, $F(1, 97) = 3.65$, $p = .059$, $\eta_p^2 = .04$. Pre-task engagement levels were similar in low demand ($M = 21.75$) and high demand ($M = 21.32$) groups, but following task performance engagement decreased under low demand ($M = 20.29$, $d = -.22$) but increased slightly under high demand ($M = 22.16$, $d = .16$). There was another significant interaction between pre-/post-task and task demand for distress, $F(1, 97) = 7.81$, $p < .01$, $\eta_p^2 = .07$. In the high task demand condition, participants reported greater distress ($M = 10.80$) after task exposure, compared to the pre-task baseline ($M = 9.22$, $d = .30$). In the low task demand condition, participants were less distressed after task exposure ($M = 6.90$, $d = -.23$) than the pre-task baseline ($M = 8.14$). Regarding worry, a significant main effect for pre-/post-task was found, $F(1, 97) = 46.14$, $p < .01$, $\eta_p^2 = .32$. Worry decreased in all conditions. In low task demand condition, worry levels decreased after task exposure ($M = 10.20$), compared to the baseline ($M = 12.92$); in high task demand condition, worry levels decreased after task exposure ($M = 10.34$), compared to the baseline ($M = 14.18$). The decrease of worry was greater in high task demand condition ($d = -.70$) than in low task demand condition ($d = -.60$).

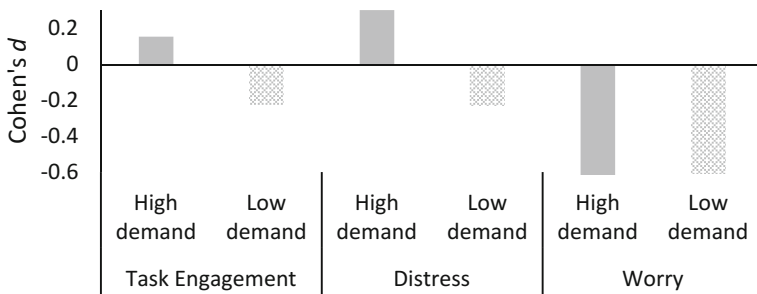


Fig. 3. Effect sizes for stress state changes (pre-task to post-task) in low/high task demand conditions.

3.3 Task Performance

A series of $2 \times 2 \times 2$ (LOA \times task demand \times task type) mixed-model ANOVAs were computed to assess the UAS operation performance, in terms of accuracy, dependence on automation, and neglect, under different demand factors.

Accuracy

Participants performed less accurately in the Weapon Release authorization task ($M = 75.7$) than in the Image Analysis task ($M = 82.3$), $F(1, 91) = 23.91$, $p < .01$, $\eta_p^2 = .21$ (see Fig. 4). Another main effect of task demand was also significant for accuracy, $F(1, 91) = 5.87$, $p < .05$, $\eta_p^2 = .06$. Participants in low task demand groups ($M = 80.9$) achieved greater accuracy than those in high task demand groups ($M = 77.1$) in the surveillance tasks. Accuracy in Weapon Release authorization task ($d = -.49$) seemed to be more vulnerable to high task demand than Image Analysis task ($d = -.28$), even though the interaction between task type and task demand was not significant.

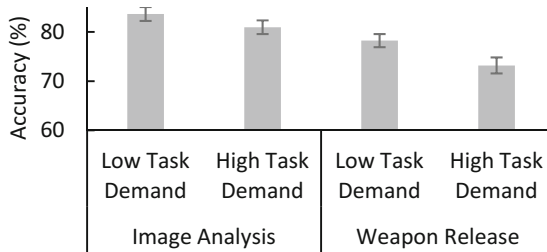


Fig. 4. Task performance (accuracy) in the Image Analysis and the Weapon Release authorization tasks for different task demand conditions. Error bars represent standard errors.

Dependence on Automation

Dependence on automation was greater in the Image Analysis task ($M = 75.6$) than in the Weapon Release authorization task ($M = 72.9$), $F(1, 91) = 5.91$, $p < .05$, $\eta_p^2 = .06$. Result revealed a significant main effect of LOA for dependence, $F(1, 91) = 5.11$, $p < .05$, $\eta_p^2 = .05$. High LOA groups ($M = 75.64$) were more dependent on automation than low LOA groups ($M = 72.76$). LOA had a stronger effect on dependence on automation in Weapon Release authorization task ($d = .46$) than in Image Analysis task ($d = .25$). Also, a near significant main effect of task demand for dependence on automation was found, $F(1, 91) = 3.92$, $p = .051$, $\eta_p^2 = .04$ (see Fig. 5). Participants showed greater dependence on automation in low task demand conditions than in high task demand conditions. In addition, the interaction between task type and task demand was also significant, $F(1, 91) = 4.76$, $p < .05$, $\eta_p^2 = .05$. In Weapon Release authorization task ($d = -.62$), task demand had a stronger effect on dependence on automation than in Image Analysis task ($d = -.03$). Specifically, in Weapon Release authorization task, participants were less dependent on automation in high task demand conditions than in low task demand conditions.

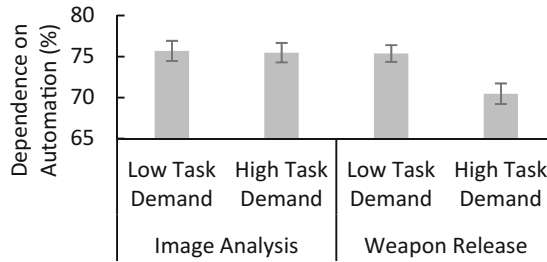


Fig. 5. Task performance (dependence on automation) in the Image Analysis and the Weapon Release authorization tasks for different task demand conditions. Error bars represent standard errors.

Neglect

Regarding neglect, there was significantly more item neglect instances in the Weapon Release authorization task ($M = 8.9$) than in the Image Analysis task ($M = 3.4$), $F(1, 91) = 94.08$, $p < .01$, $\eta_p^2 = .51$. The main effects for task demand and LOA were also significant for neglect (see Fig. 6). First, neglect was higher in high task demand groups ($M = 8.4$) than in low task demand groups ($M = 3.9$), $F(1, 91) = 19.18$, $p < .01$, $\eta_p^2 = .17$. Second, neglect was higher in high LOA conditions ($M = 7.1$) than in low LOA conditions ($M = 5.1$), $F(1, 91) = 4.20$, $p < .05$, $\eta_p^2 = .04$. Although significant, the effects of LOA manipulations on neglect were small in both Weapon Release authorization ($d = .31$) and Image Analysis ($d = .35$) tasks. In addition, the interaction between task type and task demand was significant, $F(1, 91) = 9.68$, $p < .01$, $\eta_p^2 = .10$. The effect of task demand had a stronger impact on Weapon Release authorization task ($d = .92$) than on Image Analysis task ($d = .59$). Participants in the high task demand conditions neglected the most number of items in the Weapon Release authorization task.

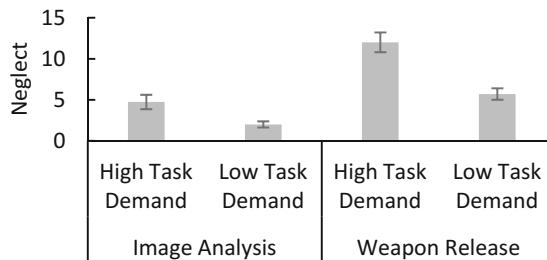


Fig. 6. Task performance (neglect) in the Image Analysis and the Weapon Release authorization tasks for different task demand conditions. Error bars represent standard errors.

4 Discussion

A main object of this study was apply a multivariate assessment strategy [20] to profile the workload, stress and performance changes associated with high task demands during simulated UAS operation. Supporting this objective, NASA-TLX, DSSQ, and performance-based measures were implemented in this study. Consistent with the SCM [12], the task demand manipulations elicited higher subjective distress and workload, as well as performance deficits. However, there was considerable variation in effect size across different indicators of overload, supporting the multivariate approach to assessing task demands impact. Other aspects of response differed qualitatively from expectation: specifically, LOAs did not affect operator workload, but affected dependence on automation.

The remainder of this section will evaluate task demand impacts on the various subjective and objective outcome measures. Most, though not all, task demand effects were consistent with prediction from the SCM [12], but variation in effect sizes indicates the limitations of relying on the capacity model alone in mitigating overload effects in the operational setting. Findings also indicated the role of participant strategy, for example, in choosing to neglect task items, and reducing automation-dependence under overload. Maladaptive strategy choices of these kinds may be addressed in training for management of overload.

4.1 Subjective Measures

Workload was assessed subjectively via NASA-TLX. Due to the convenience for administration and analysis, subjective measures are the most commonly applied method to assess workload. NASA-TLX, as one of the most popular subjective measures, assesses multiple aspects of workload. It is sensitive to overall task load and diagnostic of the nature of workload from six separate sources [20]. In this study, the level of task demand was successfully manipulated to simulate the task demand variation in UAS operations by configuring the frequency of secondary task events in the ALOA simulation. Higher subjective workload was reported by the high task demand group. According to Cohen's [31] guidelines, the 0.70 effect size for overall ratings was close to the standard for a large effect (0.80). But effect sizes varied in the individual ratings. For example, the effect sizes in physical demand, temporal demand, and frustration were medium-magnitude, whereas the effect size in mental demand was even smaller. Based on the individual ratings, physical demand was the primary source of subjective workload, presumably because frequent mouse movements were required. Therefore, designing interfaces to reduce manual demands may contribute to mitigating overall workload under high demand in UAS missions. Improvements to the interface might also reduce frustration. Time pressure is harder to mitigate because it reflects external contingencies beyond the operator's control, although future systems with greater automation of decision-making might relieve the temporal burden.

In addition, the LOA manipulations did not affect perceived workload as expected, on the basis that higher LOAs should free up attentional capacity. A possible explanation is that two intermediate levels, management-by-consent and management-by-exception, were selected from the LOA model [8] in the study. These two levels may

be too close to make a profound difference in the effect of LOA on workload. Alternatively, at the higher LOA, the operator may have reallocated attention to additional activities, such as secondary tasks, so that workload remained constant.

Subjective states of task stress were assessed via the DSSQ. This scale is widely used for profiling state change induced by task and environmental stressors in basic and applied performance tasks [24]. Using the DSSQ makes it possible to compare state change profiles across studies in the same domain. Consistent with previous studies of simulated unmanned system operations, high task demand produced greater distress, i.e., negative emotions and loss of confidence in performance [21, 22, 30]. Elevated task engagement was also reported in the higher task demand conditions, implying that the greater workload may have helped operators maintain motivation and alertness.

However, the changes in subjective state were only of small effect size, indicate that individuals can cope effectively despite high workload. Distress is normally associated with higher workload [25], but the change in distress produced by high task demands here was associated with an effect size of only .30. By contrast, Abich et al. [22] found a dual-tasking effect size of .97 in a UGV simulation study. In Abich et al.'s [22] study, participants performed surveillance tasks only, whereas ALOA requires participants to multi-task a more diverse collection of subtasks, which may have provided a more interesting and challenging assignment. Support from automation in ALOA, not present in Abich et al.'s [22] tasks, may also have limited distress. Therefore, in MAC stress intrinsic to task demands may be a minor concern operationally, although no attempt was made to simulate additional stress factors that may be present in real operations [11].

Worry was reduced relative to baseline in both task demand conditions. Typically, demanding tasks can induce decreases in worry, as attention is refocused from internal concerns to external demands [25]. The present result was also consistent with the trend of greater declines in worry in high event rate vigilance tasks [32]. By contrast, low workload, monotonous UAS tasks may lead to mind-wandering, which may, in turn, contribute to increases in worry in long-duration missions [28].

4.2 Performance-Based Measures

Three performance metrics, including accuracy, neglect, and dependence on automation, were developed for the primary surveillance tasks. Measuring task performance, such as accuracy and neglect, can provide an indication of workload and stress, especially when the operator is overloaded and performance is impaired. Performance can be impaired when demands exceed the operator's limited attentional resources [12]. Dependence on automation is an indicator of operator trust in automation. It helps to profile how operators utilize automation in coping with overloaded situations and how trust is impacted by overload.

Generally, task demand manipulations impacted performance as expected, except some findings regarding dependence on automation. Overall, larger effect sizes were found for the Weapon Release task across all three performance metrics, compared with Image Analysis. From the perspective of the SCM [12], the Weapon Release task was more demanding and required more attentional resources. Therefore, it was more vulnerable to high task demands. For both tasks, effect sizes tended to be larger for

neglect than for accuracy, implying that participants may have tried to compensate for demands by reducing the number of images they had to process, especially for Weapon Release.

In addition, dependence on automation in the Image Analysis task was consistent across task demand conditions, while significantly less dependence on automation in the Weapon Release task was observed in the high task demand condition. The moderate-magnitude effect of task demand in the Weapon Release task was unexpected. An attentional capacity perspective [29] would suggest that automation dependence should increase under higher task demands as the operator's attention is increasingly taxed. High task demands may not only deplete the limited cognitive resource, but impair operators' trust in automation. Such effects may also reflect operators' tendency to adopt task-focused coping as a strategy, relying on personal agency rather than automation for dealing with the overloaded situation. In fact, this strategy seems counterproductive, and training for operators might focus on the need to trust automation in overload situations.

Even though no effect of LOA on subjective workload and stress states was found, differences in performance-based measures were observed between the two LOA configurations. Greater dependence on automation and more neglect were observed in higher LOA conditions (management-by-exception). Higher LOA may lead to a loss of situation awareness associated with vigilance decrement and complacency issues [33] and may, in turn, result in the observed greater dependence on automation and more neglect. No significant difference in task accuracy was found between LOA conditions. This finding may suggest that considering the automation is relatively reliable, LOAs only have a subtle effect on the overall accuracy even though higher LOAs encourage operators to rely on the automation more. Also, the two LOAs were at intermediate levels close to each other. Future study may test the trend at other LOAs.

4.3 Limitations and Future Work

One limitation is the use of novice participants rather than trained operators. Applying the SCM to expertise suggests that increased skill will reduce capacity demands, and hence the vulnerability of the operator to both cognitive overload and stress [12]. However, even skilled operators remain vulnerable to overload in some circumstances, and multivariate assessment of response remains a useful methodology for guiding mitigation strategies [24]. Nevertheless, it would be desirable to further examine the roles of practice and expertise in moderating overload impacts.

Also, subjective measures are highly applicable to assessing an operator's workload when interacting with modern technologies that aid judgment and decision making [34], such as in UAS domains, and are very useful to assess stress states for testing predictions from theory [24]. But subjective measures alone are inadequate to effectively characterize workload because they can become insensitive to changes in task demand [35]. In addition, subjective measures are not applicable for diagnostic monitoring during a mission. Therefore psychophysiology assessments are also necessary for better understanding, predicting, and improving human performance in complex human-machine systems, such as UAS operations. Psychophysiology measures, such as electrocardiogram (ECG), electroencephalogram (EEG), transcranial Doppler

sonography (TCD), functional near-infrared spectroscopy (fNIR), and eye tracking, are particularly useful in assessing workload and stress state and are applicable for operational diagnostic monitoring [36].

Acknowledgement. This research was sponsored by AFOSR A9550-13-1-0016 and 13RH05COR. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of AFOSR or the US Government.

References

1. Calhoun, G.L., Ruff, H.A., Draper, M.H., Wright, E.J.: Automation-level transference effects in simulated multiple unmanned aerial vehicle control. *J. Cogn. Eng. Decis. Making* **5**(1), 55–82 (2011)
2. Guznov, S., Matthews, G., Funke, G., Dukes, A.: Use of the RoboFlag synthetic task environment to investigate workload and stress responses in UAV operation. *Behav. Res. Methods* **43**(3), 771–780 (2011)
3. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. *Hum. Factors* **46**(1), 50–80 (2004)
4. Hancock, P.A., Warm, J.S.: A dynamic model of stress and sustained attention. *Hum. Factors* **31**(5), 519–537 (1989)
5. Dixon, S.R., Wickens, C.D., Chang, D.: Mission control of multiple unmanned aerial vehicles: a workload analysis. *Human Factors* **47**(3), 479–487 (2005)
6. Hart, S.G., Wickens, C.D.: Cognitive workload in NASA human integration design handbook (NASA/SP-2010-3407). NASA, Washington, DC (2010)
7. Lee, J.D.: Affect, attention, and automation. In: Kramer, A., Wiegmann, D., Kirlik, A. (eds.) *Attention: From Theory to Practice*, pp. 73–89. Oxford University Press, New York (2006)
8. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Cybern. Part A Syst. Hum.* **30**(3), 286–297 (2000)
9. Lewis, M.: Human interaction with multiple remote robots. *Rev. Hum. Factors Ergon.* **9**(1), 121–174 (2013)
10. Matthews, G., Campbell, S.E., Falconer, S., Joyner, L.A., Huggins, J., Gilliland, K., Grier, R., Warm, J.: Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion* **2**(4), 315–340 (2002)
11. Ouma, J.A., Chappelle, W.L., Salinas, A.: Facets of occupational burnout among U.S. Air Force active duty and national guard/reserve MQ-1 Predator and MQ-9 Reaper operators (AFRL-SA-WP-TR-2011-0003). School of Aerospace Medicine Wright Patterson AFB OH (2011)
12. Matthews, G., Wohleber, R.W., Lin, J.: Stress, skilled performance, and expertise: Overload and beyond. In: Ward, P., Schraagen, J.M., Gore, J., Roth, E. (eds.) *The Oxford Handbook of Expertise*. Oxford University Press, New York (in press)
13. Wohleber, Ryan W., Matthews, G., Funke, Gregory J., Lin, J.: Considerations in physiological metric selection for online detection of operator state: a case study. In: Schmorrow, Dylan D.D., Fidopiastis, Cali M.M. (eds.) *AC 2016. LNCS (LNAI)*, vol. 9743, pp. 428–439. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39955-3_40
14. Warm, J.S., Parasuraman, R., Matthews, G.: Vigilance requires hard mental work and is stressful. *Hum. Factors* **50**(3), 433–441 (2008)

15. Warm, J.S., Matthews, G., Finomore, V.S.: Workload and stress in sustained attention. In: Hancock, P.A., Szalma, J.L. (eds.) *Performance Under Stress*, pp. 115–141. Ashgate Publishing, Aldershot (2008)
16. Raby, M., Wickens, C.D.: Strategic workload management and decision biases in aviation. *Int. J. Aviat. Psychol.* **4**(3), 211–240 (1994)
17. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* **39**(2), 230–253 (1997)
18. Dixon, S.R., Wickens, C.D., McCarley, J.S.: On the independence of compliance and reliance: Are automation false alarms worse than misses? *Hum. Factors* **49**(4), 564–572 (2007)
19. Barg-Walkow, L.H., Rogers, W.A.: The effect of incorrect reliability information on expectations, perceptions, and use of automation. *Hum. Factors* **58**(2), 242–260 (2016)
20. Matthews, G., Reinerman-Jones, L.E.: Workload assessment: How to diagnose workload issues and enhance performance. In: *Human Factors and Ergonomics Society*, Santa Monica, CA (2017)
21. Panganiban, A.R., Matthews, G.: Executive functioning protects against stress in UAV simulation. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **58**(1), 994–998 (2014)
22. Abich IV, J., Reinerman-Jones, L.E., Matthews, G.: Impact of three workload factors on simulated unmanned system intelligence, surveillance, and reconnaissance operations. *Ergonomics* **60**(6), 791–809 (2017)
23. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) *Advances in Psychology*, North-Holland, Amsterdam, The Netherlands, vol. 52, pp. 139–183 (1988)
24. Matthews, G.: Multidimensional profiling of task stress states for human factors: a brief review. *Hum. Factors* **58**(6), 801–813 (2016)
25. Matthews, G., Szalma, J.L., Panganiban, A.R., Neubauer, C., Warm, J.S.: Profiling task stress with the Dundee Stress State Questionnaire. In: Cavalcanti, L., Azevedo, S. (eds.) *Psychology of Stress: New Research*, pp. 49–90. Nova, Hauppauge (2013)
26. Matthews, G., Warm, J.S., Reinerman-Jones, L.E., Langheim, L.K., Washburn, D.A., Tripp, L.: Task engagement, cerebral blood flow velocity, and diagnostic monitoring for sustained attention. *J. Exp. Psychol. Appl.* **16**(2), 187–203 (2010)
27. Guznov, S., Matthews, G., Warm, J.S., Pfahler, M.: Training techniques for visual search in complex task environments. *Hum. Factors* **59**(7), 1139–1152 (2017)
28. Cummings, M.L., Mastracchio, C., Thornburg, K.M., Mkrtchyan, A.: Boredom and distraction in multiple unmanned vehicle supervisory control. *Interact. Comput.* **25**(1), 34–47 (2013)
29. Parasuraman, R., Manzey, D.H.: Complacency and bias in human use of automation: an attentional integration. *Hum. Factors* **52**(3), 381–410 (2010)
30. Wohleber, R.W., Matthews, G., Reinerman-Jones, L.E., Panganiban, A.R., Scribner, D.: Individual differences in resilience and affective response during simulated UAV operations. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **59**(1), 751–755 (2015)
31. Cohen, J.: A power primer. *Psychol. Bull.* **112**(1), 155–159 (1988)
32. Shaw, T.H., Matthews, G., Warm, J.S., Finomore, V.S., Silverman, L., Costa Jr., P.T.: Individual differences in vigilance: personality, ability and states of stress. *J. Res. Pers.* **44**(3), 297–308 (2010)
33. Endsley, M.R., Kiris, E.O.: The out-of-the-loop performance problem and level of control in automation. *Hum. Factors* **37**(2), 381–394 (1995)
34. Cain, B.: A review of the mental workload literature. Technical report, Defense Research and Development Toronto, Canada (2007)

35. Wierwille, W.W.: Important remaining issues in mental workload estimation. In: Hancock, P.A., Meshkati, N. (eds.) *Human Mental Workload*, pp. 315–327. North-Holland, Amsterdam (1988)
36. Matthews, G., Reinerman-Jones, L.E., Barber, D.J., Abich IV, J.: The psychometrics of mental workload: multiple measures are sensitive but divergent. *Hum. Factors* **57**(1), 125–143 (2015)



Trust in Sensing Technologies and Human Wingmen: Analogies for Human-Machine Teams

Joseph B. Lyons¹(✉), Nhut T. Ho², Lauren C. Hoffmann²,
Garrett G. Sadler², Anna Lee Van Abel¹, and Mark Wilkins³

¹ Air Force Research Laboratory, WPAFB, USA

{joseph.lyons.6,anna.van_abel}@us.af.mil

² NVH Human Systems Integration, LLC, Los Angeles, USA

nhut.ho.51@gmail.com, lauren.c.hoffmann@gmail.com,
garrett.g.sadler@gmail.com

³ Office of the Secretary of Defense, Arlington, USA

mark.a.wilkins10.ctr@mail.mil

Abstract. The true value of a human-machine team (HMT) consisting of a capable human and an automated or autonomous system will depend, in part, on the richness and dynamic nature of the interactions and degree of shared awareness between the human and the technology. Contemporary views of HMTs emphasize the notion of bidirectional transparency, one type of which is Robot-of-Human (RoH) transparency. Technologies that are capable of RoH transparency may have awareness of human physiological and cognitive states, and adapt their behavior based on these states thus providing augmentation to operators. Yet despite the burgeoning presence of health monitoring devices, little is known about how humans feel about an automated system using sensing capabilities to augment them in a work environment. The current study provides some preliminary data on user acceptance of sensing capabilities on automated systems. The present research examines an emerging predictor of trust in automation, Perfect Automation Schema, as a predictor of trust in the sensing capabilities. Additionally, the current study examines trust of a human wingman as an analogy for looking at trust within the context of a HMT. The findings suggest that Perfect Automation Schema is related to some facets of sensing technology acceptance. Further, trust of a human wingman is contingent on familiarity and experience.

Keywords: Trust in automation · Autonomy · Human-machine teaming
Military

1 Introduction

Advances in modern technology place humans in contexts where machines may someday be partners versus tools [24]. To achieve this vision, machines will need to engage in team-based behaviors in collaboration with humans. Some of these team-based behaviors may involve monitoring human physiological activity and task

performance. A good teammate, after all, is aware of when her/his teammates are stressed, overloaded, or just not engaged. Team members often use this information to support or “back-up” another team member. This back-up behavior exemplifies being a good teammate [1]. Understanding and acting on degraded human performance (either physiological or task-based) could be a way for advanced technology to augment human performance in military environments, yet little is known about how such technologies would be accepted or rejected among military operators.

Modern society has witnessed an explosion of devices for monitoring health, activity level, and other factors. Yet, the bulk of these tools are voluntary and while useful for tracking fitness, many of these tools are entertainment-centric and are not tied to an augmentation strategy from the technology. In other words, these tools provide information only and it is up to the human to utilize their guidance in most cases. From a levels-of-automation standpoint, this would constitute information acquisition and analysis [19], which is on the lower end of the levels of automation spectrum. What happens when these systems begin to integrate information about humans’ states into their decision processes? Furthermore, what happens when these systems are granted authority to redirect their actions based on an understanding of human states and one’s performance threshold? Imagine a world where one’s watch can dictate whether or not a driver is alert enough to drive, or where one’s fitness monitor prohibits the purchase of a desired tasty treat. With the advent of novel technologies desired to sense and augment humans, researchers must consider human acceptance, or trust, of the technologies and their behavior.

Trust refers to one’s willingness to be vulnerable to another entity [14]. Recent literature has reviewed the construct of trust as it relates to trust of machines [9, 22]. Much of this literature has examined the effects of reliability, performance, and error types on trust and other outcomes [9, 18, 21, 22]. But an emerging trend within this literature focuses on concepts such as transparency, i.e., methods for establishing shared awareness and shared intent [see 10]. Transparency manipulations have been shown to influence trust of automated systems [8, 13, 15]. Most of the transparency-based designs examined in prior research use interface-based features to convey information about the real-time activities associated with an automated tool, or they might use an interface to display the rationale for an automation’s decision or recommendation. In all such cases, these are examples of Robot-to-Human (RtH) transparency as discussed by [11] wherein the robot (or automation in this case) communicates task-based and analytical awareness information to the human in an effort to foster greater shared awareness. Of the various transparency facets, Robot-of-Human (RoH) transparency refers to when a system uses information about the human’s state to guide its interaction with the human and to explain its behavior. Knowledge of human workload, stress, boredom, or degraded physiological capacity could be instrumental in determining when a system should intervene in a human operator’s task. The awareness of human states and a system’s augmentation in relation to those states has been examined in the literature on adaptive automation.

Adaptive automation is automation that can invoke a higher or lower level of automation based on an operator’s state in critical situations – such as safety critical situations [2]. It is believed that adaptive automation can reduce human-automation interaction errors [2, 6]. Research by [5] examined a form of automation that used

human Electroencephalography (EEG) signals as a means to understand the human's cognitive workload and to, in turn, determine the appropriate time to interrupt the human operator without overloading her/him. They found performance benefits for the system under high workload conditions [5]. Adaptive systems that trigger based on human performance decrements have been present in the automotive community for years. Such systems may engage in augmentation strategies such as (1) arousing a driver's attention to encourage greater attention allocation to potential risks, (2) providing warnings that encourage the driver to make appropriate decisions and actions to avoid accidents, and (3) using fully-automated control systems to take action when no action by the human is detected and an action is needed to avoid an accident [10]. The military too, has recently fielded a fully-automated safety system that will assume control of certain aircraft (e.g., F-16 fighters) to prevent ground collision [8]. These are examples of adaptive systems that monitor human performance thresholds. Yet, little is known about how humans view systems that are capable of sensing our physiology and altering their actions based on that understanding. The current research investigates several possible sensing capabilities and gauges operator acceptance of these methods. The current study also examines the Perfect Automation Schema (PAS) as the predictor of these attitudes.

The construct of PAS has recently gained attention as a trust antecedent. PAS has been conceptualized as a two-factor construct consisting of High Expectations (HE) and All-or-None beliefs (AoN) [16]. People with higher HE believe that automated systems are highly reliable, whereas those with AoN beliefs feel that any faults on the part of the automation means that the whole system is broken. Research has shown that the HE and AoN facets of PAS do influence trust perceptions, however the studies have shown inconsistent results in terms of whether it is HE or AoN that is related to trust perceptions [16, 20]. As such, PAS was included to examine if and how it is related to acceptance of sensing technologies.

The whole purpose of developing technologies that are capable of sensing and reacting to human states is to promote more effective teaming between the humans and machines. It is clear that autonomous systems and the human ability and preference to team with these technologies is an important part of future research doctrine, notably within the department of Defense (DoD) [3, 4]. However, there are many research challenges associated with the notion of HMT. Using machines as teammates also calls into question how such systems will be evaluated. Analogies can be drawn using human-human teaming as a comparison. There is a vast literature on human-human teams which may inform the design and evaluation of human-machine teams [24]. In the military context, the Air Force is exploring the concept of an autonomous wingman. Evaluations strategies of concepts like this need to account for the natural variance that occurs through human-human interaction when using humans as a comparison group, lest the evaluation be biased. As such, the current study examined pilot trust of different types of human wingmen.

Two factors must be considered when using human teams as a benchmark for comparison of trust for an autonomous system designed to team with humans, namely familiarity and experience. It is likely that the human team members used as a benchmark will be familiar with one another. This familiarity should positively influence trust perceptions [17, 23]. Thus, trust comparisons between human teams and

HMTs may be contaminated by human familiarity – albeit, if designed poorly. Experience, specifically task experience, should also influence trust perceptions. Experience should be associated with greater learned trust [as noted by 8] and should be associated with greater perceived ability – which is a known trust antecedent [14]. If trust of a human teammate with considerable task experience was compared to trust of an autonomous system, the system will likely be subjected to a trust-based biases favoring the human simply based on experience. Thus, the current study will investigate how familiarity and experience influence trust of a human wingman to demarcate the impact of these factors and to show potential bias that would inevitably proliferate poorly-designed human versus machine comparisons in the HMT domain.

Given that several sensing capabilities were examined in the present study, no explicit hypotheses are posited to suggest greater or less acceptance of one type over another. Rather this study described the acceptance levels across the different types. It was expected that both the HE [greater] and AoN [less] facets of PAS would be associated with acceptance of the sensing technologies. Finally, it was expected that trust of a human wingman would increase as familiarity and experience increase.

1.1 Participants

Seventy-four F-16 pilots served as the participants for this research. They averaged 1700 flight hours and each was an operational pilot versus a trainee.

1.2 Materials and Procedure

As part of a larger study on pilot trust of automated collision avoidance technologies, pilots were asked to respond to an online survey which gauged their acceptance of sensing technologies. The sensing technologies varied in focus and design intent as noted below. The pilots were also asked to respond to three items which gauged their trust of a human wingman and they completed items for measuring the Perfect Automation Schema.

1.2.1 Sensing Technologies

Using a 7-point Likert scale where 1 = strongly disagree and 7 = strongly agree, participants were asked to rate their agreement with the following items (each item was prefaced by “I would be comfortable with an automated system on my aircraft that...”): (1) monitored my heart rate, (2) monitored my brain activity, (3) assessed my task performance, (4) assessed my mental alertness, (5) changed its behavior based on an understanding of my brain activity, (6) changed its behavior based on an understanding of my task performance, (7) changed its behavior based on an understanding of my mental alertness.

1.2.2 Wingman Trust Items

Using a 7-point Likert scale where 1 = strongly disagree and 7 = strongly agree, participants were asked to rate their agreement with the following items: (1) I would trust a human wingman who was *unfamiliar and inexperienced*, (2) I would trust a

human wingman that was *unfamiliar but experienced*, and (3) I would trust a human wingman who was *familiar and experienced*.

1.2.3 Perfect Automation Schema

There were two scales related to the Perfect Automation Schema: High Expectations (HE) and All-or-None (AoN) beliefs [16]. Using a 7-point Likert scale where 1 = strongly disagree and 7 = strongly agree, participants were asked to rate their agreement with the following items: [HE items] (1) Automated systems have 100% perfect performance, (2) Automated systems rarely make mistakes, (3) Automated systems can always be counted on to make accurate decisions, (4) Automated systems make more mistakes than people realize [reverse-coded]; [AoN items] (1) If an automated system makes an error, then it is broken, (2) If an automated system makes a mistake, then it is completely useless, (3) Only faulty automated systems provide imperfect results.

2 Results

As shown in Table 1, the pilots unexpectedly reported similar comfort levels for all of the sensing items. As shown in Table 2, the Perfect Automation Schema was associated with some of the sensing items. Specifically, HE was marginally associated with greater comfort of technologies that change their behavior based on one's task performance. AoN was associated with less comfort for technologies that assess mental alertness and those that can change their behavior based on one's physiological activity. AoN was also marginally associated with less comfort for technologies that assess task performance and technologies that change their behavior based on one's mental alertness. As shown in Fig. 1, the wingman analyses followed the expected trend that trust increased as familiarity and experience increase. Trust varied based on the familiarity and experience of the wingman, $F(1, 73) = 256.39, p < .001$, and trust was lowest for unfamiliar-inexperienced wingmen and highest for familiar-experienced wingman. The differences were reliable at each increment of familiarity/experience.

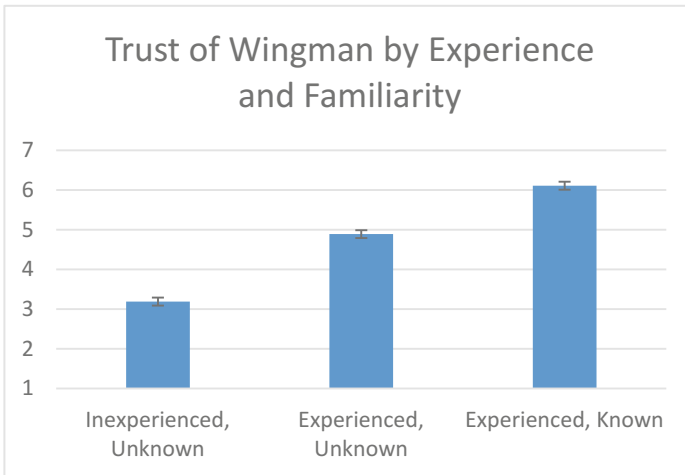
Table 1. Descriptive statistics for the sensing items

Item	Mean	Standard deviation
Monitored my heart rate	3.66	2.04
Monitored my brain activity	3.37	2.03
Assessed my task performance	3.68	1.96
Assessed my mental alertness	3.74	1.91
Changed its behavior based on an understanding of my physiological activity	3.46	1.82
Changed its behavior based on an understanding of my task performance	3.50	1.86
Changed its behavior based on an understanding of my mental alertness	3.43	1.83

Table 2. Correlations between the sensing items and High Expectations (HE) and All-on-None (AoN) beliefs.

Item	HE	AoN
Monitored my heart rate	.10	-.22 [□]
Monitored my brain activity	.09	-.18
Assessed my task performance	.07	-.20 [□]
Assessed my mental alertness	.11	-.23*
Changed its behavior based on an understanding of my physiological activity	.15	-.27*
Changed its behavior based on an understanding of my task performance	.20 [□]	-.15
Changed its behavior based on an understanding of my mental alertness	.18	-.20 [□]

Notes. [□]p < .1. *p < .05.

**Fig. 1.** Wingman trust by familiarity and experience.

3 Discussion and Implications

Teaming relationships with advanced technology is complicated and success in these relationships will be predicated on the ability of the humans and machines to establish shared awareness and shared intent. Contemporary researchers have suggested that bidirectional transparency may be one method to help foster shared awareness between humans and machine partners [11]. This will require that machines ingest and integrate information about human states (physiological and psychological) into their actions. The current study examined human acceptance of general sensing technologies that varied in their input (i.e., physiological metrics - such as heart rate or neurological signals, task performance, or mental alertness) and their targeted response (e.g., merely assessment versus augmentation). The results showed that operational pilots did not favor (or disfavor) one sensing type over another. Pilots evidenced moderate comfort

levels for sensing technologies that spanned the gamut of capabilities from sensing heart rate and neurological activity, assessing task performance and mental alertness, and taking action based on the sensed signals. The pilots did not seem to differentiate between technologies that simply sensed, versus those that sensed and augmented the human based on the sensed information.

It is possible that pilots are getting more accustomed to technologies that sense and augment them in operations. The Air Force recently fielded the Automatic Ground Collision Avoidance System (AGCAS) which senses the pilot's flight performance and automatically recovers the aircraft when a collision is detected. Despite the fully-automated nature of AGCAS, pilots have grown to trust the system and have accepted it as a useful safety technology [8]. Thus, it is possible that operators such as fighter pilots, are getting more accepting of sensing and augmentation technologies in general. However, an alternative explanation is that the nature of the technologies described in the current study were too high-level to warrant resistance. In all cases, the technologies examined in this study were described absent details on how the system would sense (i.e., what sorts of sensors would be used), how the data would be used, and what implications of the augmentation would be for the pilots. Many pilots may be comfortable with commercial products that gauge physiology for instance, yet these same pilots may report resistance to wearing a cumbersome set of electrodes under their flight helmet due to the physical discomfort and the lack of familiarity with such systems. None of the technologies discussed in the current study included details about how they would be used in the cockpit, and this lack of detail may have promoted more innocuous perceptions of the technologies in general. The intended use of the data is also very relevant. If pilots were to think that the sensing technologies could be used punitively or that they may result in flight disqualification, then certainly the technologies would be faced with greater resistance. Finally, there were no details about how the augmentation would occur. Specific details about how the system would engage in augmentation may be subject to greater resistance than the general ideas of augmentation. While the current study sought to understand acceptance of sensing and augmentation technologies in general, the lack of details associated with the technologies may have masked potential resistance among pilots. Future research should examine acceptance of specific sensing and augmentation technologies. One such study examined pilot trust of an automated air collision avoidance system and it noted several trust barriers by pilots for this technology [12].

In terms of a trust antecedent, PAS appeared to be related to acceptance of the technologies. Specifically, AoN beliefs were related to less acceptance of the sensing and augmentation technologies. The tendency to have AoN beliefs are associated with individual perceptions that advanced technologies always useful or always ineffective. The dichotomous view of technology exemplified by AoN beliefs could be a useful predictor of operator acceptance and/or rejection of novel technologies. Surprisingly, HE was not associated with acceptance of the technologies, albeit with the exception of a marginal positive relationship with acceptance of technologies that augment based on one's task performance. The present findings add to a growing literature on individual differences of the trustor that are associated with trust in automation [16, 20]. Variability in trust can also be based on features of the trustee.

Trust of a teammate can be based on a number of factors. The current study examined how familiarity and perceived task experience influence trust. As expected, trust of a wingman increased with greater familiarity and higher task experience. While this finding is not surprising, it raises an important point that relates to HMTs. Test plans that seek to compare the effectiveness of a HMT may use human-human teams as an analogous comparison, and this comparison is an understandable benchmark. HMTs should be at least as effective as their human counterparts, right? Well, maybe... However, comparisons to human teammates can artificially bias the evaluations in favor of the human teams if not properly designed. In particular, it is likely that human-human teammates will have a *de facto* benefit for trust by virtue of their increased familiarity in comparison to an unfamiliar machine under test. In this case, factors such as perceived task experience and familiarity must be accounted for in the comparisons. This accounting, however, is easier said than done. Specifically, to develop task experience and familiarity, there exists a need for a teaming agent that facilitates teamwork between the human and the machine by adapting to the preferences for interaction of each partner (human or machine). At the same time, it is important that this facilitated interaction is transparent and bi-directional (i.e., comprising of RtH and RoH), which can be achieved by interactive “training” in which both the human and machine learn about each other while taking each partner’s preferences and strengths and weaknesses into consideration. A seminal research effort for building this type of teaming agent has been spearheaded for the NASA Reduced Crew Operations Program [7], but further research is needed to conceptualize and build an agent that can be generalized for any application.

The current study has a number of implications. First, pilots reported moderate levels of acceptance to general technologies that seek to sense and augment in various ways. Researchers should continue to gauge pilot comfort levels with novel technologies to avoid fielding a new tool that will be rejected by the operators. According to the present data, pilots did not seem bothered by higher-level automated systems that not only sense but also augment them. Care needs to be taken that future sensing and augmentation systems are not resisted based on lack of trust. Engineers and researchers should consider pilot preferences for sensor placement and feasibility as pilots may show significant resistance to sensors that are painful, distracting, and disliked. Further, the use and implications of use for the technologies need to be considered. Technologies that carry the potential for punitive action and or those that have the potential to impact a pilot’s flight readiness may be faced with resistance. The technologies explored in the current study may have lacked the specific details to have revealed these nuisances.

Operator individual differences such as the AoN component of the PAS could be useful predictors of resistance to novel technologies. Engineers who seek to field new technologies need to be aware that individuals may naturally vary in their acceptance of new technologies. Yet, many of these individual differences can be assessed and used to identify individuals who may be more resistance to the technologies. If individuals are believed to be resistant, care must be taken to avoid overselling unreliable tools. In contrast, designers should use transparency guidelines to promote shared awareness and shared intent between humans and new technologies [2, 8, 11, 13, 15].

Finally, when evaluating trust of technologies in the context of a HMT, researchers should be careful with using human teams as a comparative benchmark. Poorly designed comparisons between an unfamiliar technology as a teammate compared teams of humans that have experience working together will introduce an unfair bias against the technologies. Comparisons might be made in two ways: (1) with human teams who have no experience working together and with humans who are not familiar with one another; or (2) with human-machine teams that have an agent whose role is to facilitate teamwork by adapting each partner's preferences for interaction. This will create an even playing field between the HMT and the human teams.

References

1. Burke, C.S., Stagl, K.C., Salas, E., Pierce, L., Kendall, D.: Understanding team adaptation: a conceptual analysis and model. *J. Appl. Psychol.* **91**(6), 1189–1207 (2006)
2. Chen, J.Y.C., Barnes, M.J.: Human-agent teaming for multirobot control: a review of the human factors issues. *IEEE Trans. Hum. Mach. Syst.* **44**, 13–29 (2014)
3. Defense Science Board (DSB) Task Force on the Role of Autonomy in Department of Defense (DoD) Systems. Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, Washington, DC (2012)
4. Defense Science Board (DSB) Summer Study on Autonomy. Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, Washington, DC (2016)
5. Dorneich, M.C., Ververs, P.M., Mathan, S., Whitlow, S., Hayes, C.C.: Considering etiquette in the design of an adaptive system. *J. Cogn. Eng. Decis. Making* **6**(2), 243–265 (2012)
6. Hancock, P.A., Jagacinski, R.J., Parasuraman, R., Wickens, C.D., Wilson, G.F., Kaber, D. B.: Human-automation interaction research: past, present, and future. *Ergon. Des.* **21**(9), 9–14 (2013)
7. Ho, N., Johnson, W., Lachter, J., Brandt, S., Panesar, K., Wakeland, K., Sadler, G., Wilson, N., Nguyen, B., Shively, R.: Application of human-autonomy teaming to an advanced ground station for reduced crew operations. In: 36th Digital Avionics Systems Conference, St. Petersburg, Florida, USA, 17–21 September 2017
8. Ho, N.T., Sadler, G.G., Hoffmann, L.C., Lyons, J.B., Ferguson, W.E., Wilkins, M.: A longitudinal field study of auto-GCAS acceptance and trust: first year results and implications. *J. Cogn. Eng. Decis. Making* **11**, 239–251 (2017)
9. Hoff, K.A., Bashir, M.: Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Fact.* **57**, 407–434 (2015)
10. Inagaki, T.: Smart collaboration between humans and machines based on mutual understanding. *Annu. Rev. Control* **32**, 253–261 (2008)
11. Lyons, J.B.: Being transparent about transparency: A model for human-robot interaction. In: Sofge, D., Kruijff, G.J., Lawless, W.F. (eds.) *Trust and Autonomous Systems: Papers from the AAAI Spring Symposium (Technical Report SS-13-07)*. AAAI Press, Menlo Park (2013)
12. Lyons, J.B., Ho, N.T., Van Abel, A.L., Hoffmann, L.C., Eric Ferguson, W., Sadler, G.G., Grigsby, M.A., Burns, A.C.: Exploring trust barriers to future autonomy: a qualitative look. In: Cassenti, D.N. (ed.) *AHFE 2017. AISC*, vol. 591, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60591-3_1
13. Lyons, J.B., Koltai, K.S., Ho, N.T., Johnson, W.B., Smith, D.E., Shively, J.R.: Engineering trust in complex automated systems. *Ergon. Des.* **24**, 13–17 (2016)

14. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrated model of organizational trust. *Acad. Manag. Rev.* **20**, 709–734 (1995)
15. Mercado, J.E., Rupp, M.A., Chen, J.Y.C., Barnes, M.J., Barber, D., Procci, K.: Intelligent agent transparency in human-agent teaming for multi-UxV management. *Hum. Fact.* **58**(3), 401–415 (2016)
16. Merritt, S.M., Unnerstall, J.L., Lee, D., Huber, K.: Measuring individual differences in the perfect automation schema. *Hum. Fact.* **57**, 740–753 (2015)
17. Moss, S.A., Garivaldis, F.J., Toukhsati, S.R.: The perceived similarity of other individuals: the contaminating effects of familiarity and neuroticism. *Pers. Individ. Differ.* **43**, 401–412 (2007)
18. Onnasch, L., Wickens, C.D., Li, H., Manzey, D.: Human performance consequences of stages and levels of automation: an integrated meta-analysis. *Hum. Fact.* **56**, 476–488 (2014)
19. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **30**, 573–583 (2000)
20. Pop, V.L., Shrewsbury, A., Durso, F.T.: Individual differences in the calibration of trust in automation. *Hum. Fact.* **57**, 545–556 (2015)
21. Rice, S.: Examining single and multiple-process theories of trust in automation. *J. Gen. Psychol.* **136**, 303–319 (2009)
22. Schaefer, K.E., Chen, J.Y., Szalma, J.L., Hancock, P.A.: A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Hum. Fact.* **58**(3), 377–400 (2016)
23. Webber, S.S.: Development of cognitive and affective trust in teams. *Small Group Res.* **39**(6), 746–769 (2008)
24. Wynne, K.T., Lyons, J.B.: An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science* (in press)



Deep Convolutional Neural Networks and Power Spectral Density Features for Motor Imagery Classification of EEG Signals

A. F. Pérez-Zapata¹, A. F. Cardona-Escobar¹, J. A. Jaramillo-Garzón²,
and Gloria M. Díaz¹(✉)

¹ Instituto Tecnológico Metropolitano, Medellín, Colombia
{andresperez75267, andrescardona134713}@correo.itm.edu.co,
gloriadiaz@itm.edu.co

² Universidad de Caldas, Manizales, Colombia
jorge.jaramillo@ucaldas.edu.co

Abstract. A Brain-Computer Interface (BCI) is a communication and control system that attempts to provide real-time interaction between a user and a computer device, based on the brain electrical signals that are generated when user imagine specific movements or actions. For doing so, classification models are developed to identify the user movement intention according to specific signal features. This paper presents a classification model to BCI that is based on the processing of Electroencephalography (EEG) signals. The power spectral density (PSD) representation of EEG signals is used for training a deep Convolutional Neural Network (CNN) that is able to differentiate among four different movement intentions: left-hand movement, right-hand movement, feet movement, and tongue movement. Performance evaluation results reported a mean accuracy of 0.8797 ± 0.0296 for the well-known BCI Competition IV Dataset 2a, which outperform state-of-the-art approaches.

Keywords: Brain Computer Interface (BCI)
Electroencephalography (EEG) · Power Spectral Density (PSD)
Deep learning · Convolutional Neural Network (CNN)

1 Introduction

A Brain-Computer Interface (BCI) is a communication and control system that measures and analyzes brain activity from biological signals to provide real-time interaction between a human and computer devices [1]. A BCI system is commonly composed by four main stages: signal acquisition, preprocessing, feature extraction and representation, and movement intention identification [2]. Acquisition of neurophysiological signals can be performed using different non-invasive and invasive methodologies. However, as can be expected, the invasive

approaches come with risks of causing physical injury to the individual [3]; for this reason, non-invasive systems such as EEG are a more convenient alternative to preserve patient health, since they are based on the superficial placement of electrodes on the scalp, avoiding cortical implants and surgeries that endanger subject integrity. BCI systems based on EEG can use different types of signals from brain electrical activity such as Slow Cortical Potentials (SCP), Visual Evoked Potentials (VEP) and Sensorimotor Rhythms from Motor Imagery (MI) [4]. MI is the process of imagine certain action or movements without performing the action or movement itself. It is known that MI process activates approximately the same brain regions that the real movement would do [5]. The second stage, feature extraction and representation, concerns to the extraction of the most relevant characteristics of the acquired neurophysiological signals and representing it in a feature vector. This step is very relevant to non-invasive acquisition methodologies since they are very noise-prone [6]. The third stage is the identification of the desired action from the acquired signals for its later execution on a device, a task known as signal classification.

This work proposes the use of Power Spectral Density (PSD) and deep learning techniques to develop an EEG signal classification model for MI. PSD estimation method identifies dominant frequencies that allow to find good separability patterns in the EEG signals, in this work two well-known PSD estimation functions were evaluated i.e. Welch and Periodogram. PSD features are then learned by a Convolutional Neural Network in order to differentiate among four movement intentions i.e. left-hand movement, right-hand movement, feet movement, and tongue movement. As BCI signals are not well suited for CNN due to the lack of spatial correlation among channels, a window-based preprocessing approach was implemented, which reorganizes the position of channels according to their distance in the scalp. Additionally, a sliding-window scheme was performed to reach the data required to train a deep model.

This paper is organized as follows: in the next section, a brief summary of the related works described in the literature is presented. Section 3 introduces the technical details of the proposed approach, describing the acquisition (dataset description), information representation based on PSD estimation, and CNN based classification stages. Section 4 presents experimental results that show the reliability of the proposed architecture, and finally, Sect. 5 presents the conclusions and discusses perspectives of future works.

2 Related Works

Development of techniques for identifying MI from EEG signals has been a strong field of research in recent decades [7]. In the pre-processing stage, the development of techniques to reduce noise is sought. At this point, techniques as Common Spatial Pattern (CSP), Principal Component Analysis (PCA), Common Average Reference (CAR), among others have been used [8]. In the field of feature extraction and representation, it has been attempted to develop techniques that allow identifying or finding discriminating characteristics among the

different signals. From the strategies that have been proposed, stand out those based on Wavelet Decomposition, Sub-band Energy and Shannon Entropy, among others. (A complete review of recent works at this point can be found at [9]). Finally, in the field of signal classification, there are many techniques that have been proposed, including k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Artificial Neural Networks [10] and recently Deep Neural Networks [11].

Several datasets have been published to promote research on BCI systems based on EEG. Among them, the BCI Competition IV dataset 2a [12] have been widely used in the literature and will be used for evaluating the proposed approach. Thus, at the following, the best performing approaches to this database will be described in order to establish a comparative context of the results.

In 2015, Bashashati et al. [13] presented a comprehensive comparison of classification models for identifying movement intentions in several EEG datasets, including the BCI Competition IV - 2a. The methodology used starts with a filter bank composed by fifth order Butterworth band pass filters array. Then, a spatial filter (common spatial pattern) is applied before extracting signal features that were then used for evaluating the classification models. In the BCI IV competition 2a dataset, the logistic regression and Multi Layer Perceptron (MLP) classifiers outperformed others, getting a mean accuracy of 74.33% and 74.42% respectively. However, a high dispersion between the accuracy of each subject makes the results unreliable.

Recently, Helal et al. [14] developed an LDA based method with Autoencoders, which was composed of five stages: pre-processing, feature extraction, dimensionality reduction, classification, and evaluation. In the preprocessing, signal artifacts were removed by applying whitening, CAR and Z-Score normalization to the raw data. Band-power method was implemented as feature extraction, PCA and Autoencoders were used for dimensionality reduction, and finally, LDA was employed in the classification task. Reported results showed that Autoencoders with non-linear activation function (Sigmoid) achieves better performance compared to PCA, getting a mean classification accuracy of 67%.

Methodologies based on deep learning with filter banks and CSP have been also proposed with promising results. Merinov et al. [15] proposed a spatial filter network (SFN). Their approach was evaluated using the BCI competition III dataset 3a and the BCI competition IV dataset 2a. In the preprocessing step, time segments between 0.5–4 s from the instruction cue on set, are taken for obtaining segments of 3.5 s as training set, then, the signals were passed through the frequency bands range of the original CSP algorithm. In the SFN, authors utilized cross-entropy loss function, a batch learning scheme and data augmentation. Each learning epoch is composed by 100 batches, in a 5-fold cross validation loop. This approach reported a best accuracy of 0.65. Sakhavi et al. [16] used the Filter-Bank CSP (FBCSP) proposed in [17]. A bank of 9 filters from 4 to 40 Hz, with a width of 4 Hz was initially applied for extracting a set of features that were then selected by a mutual information feature selection algorithm. For the CSP process, four pairs of spatial filters were picked for each

frequency band, finally the proposed parallel CNN and linear architecture were implemented to decode the final movement intention class. Mean accuracy for all subjects of 70.60% in the BCI IV Competition 2a dataset was obtained. Yang et al. [18] proposed Augmented CSP (ACSP) features based on a varying the frequency bands with different bandwidths to cover as many bands as possible, and a CNN based methodology to classify EEG signals, using the BCI IV competition 2a dataset. They proposed a way to select the feature maps, namely frequency complementary map selection (FCMS), and compared it with random map selection (RMS) and with the selection of all feature maps (SFM). Average cross-validation accuracy of 68.45% for FCMS and 69.27% for SFM was achieved. Nonetheless, approaches based on Filter banks with CSP highly depends on the filter bands selection and this can cause loss of relevant information.

3 Materials and Methods

3.1 Method Overview

Figure 1 illustrates the main stages of the proposed approach. EEG signals are initially processed to improve two data characteristics: the amount of data, performed by a data augmentation process based on a sub-window sampling, and the spatial relations between electrodes, aimed by a channel reordering. Then, a PSD based feature extraction approach is carried out for each subsampled signal. Finally, feature vectors are used to train and test a deep learning classification model in a 5-Fold cross validation strategy.

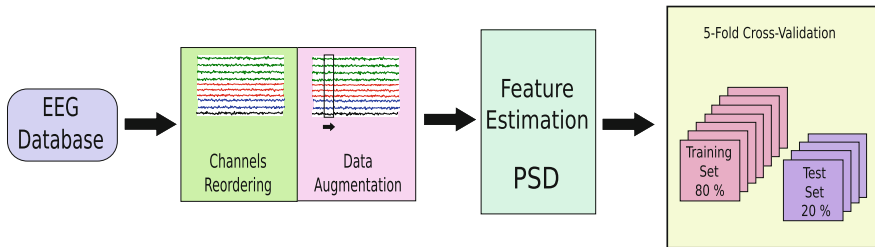


Fig. 1. Summary of the proposed approach for motor imagery classification

3.2 Dataset Description

Performance evaluation is carried out using the BCI Competition IV dataset 2a [12], which consist of EEG signals taken from 9 subjects, with a sample rate of 250 Hz. Each subject was taken four motor imagery tasks: left hand movement, right hand movement, and feet and tongue movements, which were labeled as classes 1, 2, 3 and 4 respectively. For each subject, two sessions with six runs each one were recorded. One run was composed of 48 trials (12 per class), for a total of 288 trials per session.

The dataset is originally in GDF format, which is used for biomedical signals. The signals are in an array of 672528×25 , the last three columns correspond to electrodes intended to acquire Electrooculography (EOG) signals, which were not taken into account in the classification task.

3.3 Data Preprocessing

Because CNN requires a large amount of data, and spatial relationships between them, which are not necessarily found in one-dimensional MI signals, the pre-processing stage is herein composed of two steps, the first one seeks to expand the number of training and testing data, and the second, reorganize the order of the electrodes according to its proximity into the scalp, seeking to establish a spatial correlation between the signals.

Data Augmentation Based on Sub-window Extraction. A window-based approach over temporal signals was carried out, in order to extract the enough amount of pseudo images as [19] proposed, but with few differences in the overlapping. For the sake of generality, we introduce some notation; for each subject suppose a matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$ where T is the amount of temporal observations and N is the number of channels in the recording process. The main purpose is to use a sliding window of size τ to slice the temporal axis in several overlapping windows given by $\mathbf{x}_i \in \mathbb{R}^{\tau \times N} \forall i = 1, \dots, T - \tau + 1$. In order to avoid over fitting, the overlap size was 95% (this value was selected as rule-of-thumb) instead of the maximum overlapping size that gives an almost identical pseudo image, redefining the classification task. This process can be seen on Fig. 2.

Channel Reordering. Commonly, CNNs are trained using 2D data, where spatial correlation is guaranteed. However, motor imagery signals are not spatially organized. To overcome this problem, we ensure spatial correlations among channels by applying k-means clustering over them, in line with its site in the scalp, unlike [19], where K Nearest Neighbor was applied. Site in the scalp was based on the electrode montage corresponding to the international 10-20 system used on [12], with the value of the coordinates representing the distance between

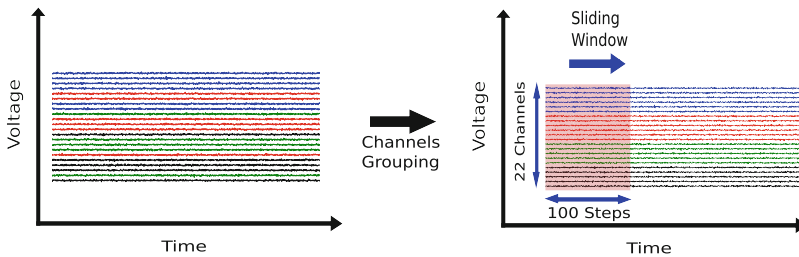


Fig. 2. Preprocessing of motor imagery signals.

the electrodes, having an inter-electrode distance of 3.5 cm, obtaining the spatial matrix:

$$SP = [0 \ 0; -7 \ -3.5; -3.5 \ -3.5; 0 \ -3.5; 3.5 \ -3.5; 7 \ -3.5; -10.5 \ -7; -7 \ -7; -3.5 \ -7; 0 \ -7; 3.5 \ -7; 7 \ -7; 10.5 \ -7; -7 \ -7 \ -10.5; -3.5 \ -10.5; 0 \ -10.5; 3.5 \ -10.5; 7 \ -10.5; -3.5 \ -14; 0 \ -14; 3.5 \ -14; 0 \ -17.5]$$

Electrode 1 takes the coordinates (0,0) and the others take its coordinates according to its distance in cms to electrode 1. Then, the 22 channels were organized in four groups, within each group channels were ranked according to the distance to the center of its group, finally, all groups are stacked into a single matrix. This method was used by Walker et al. [19].

3.4 Feature Extraction Based on Power Spectral Density

PSD gives an estimation about the power of a signal at different frequencies for any temporal signal [20].

Periodogram Function. This measure is computed over a signal to find the spectrum in different parts. The square of these results are known as periodograms, defined by the following expression [21]:

$$P = \frac{1}{N} A^2[\omega] \tag{1}$$

where $A^2[\omega]$ is the square of the FFT for a signal $a[k]$ with N observations. In this work, PSD was employed for feature extraction in EEG signals using FFT, this is done for each channel; that is, given a record $\mathbf{x}_i \in \mathbb{R}^{T \times N}$ we computed periodograms for the channel $j \ \forall j = 1, \dots, N$. PSD implementation available in the SciPY package [22] was used.

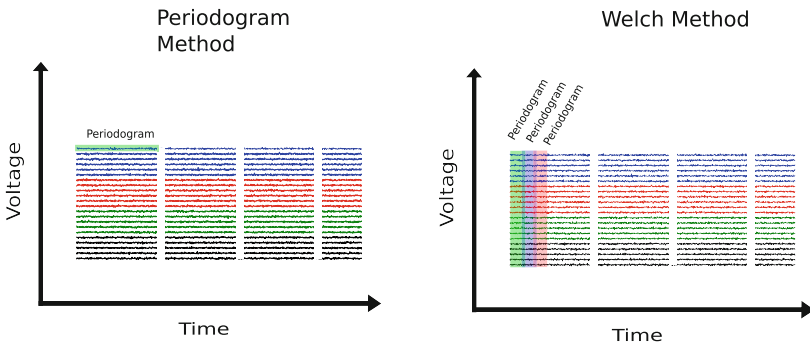


Fig. 3. PSD estimation approaches.

Welch Function. This nonparametric approach for PSD estimation is regularly employed, its main advantage is the no assumption about the distribution of data [23]. The main drawback of the periodogram estimator is the high variance [24]. Conversely, the Welch method presented in [20] is a more efficient estimator which is based on overlapping sections, where a sliding window is used to find the periodogram in those segments. Finally the result is obtained by averaging the estimations of all sections as is explained in [23]. Figure 3 shows this two PSD estimation approaches.

3.5 Deep Network Architecture

An outline of the architecture used can be seen in Fig. 4. Convolutional Neural Networks are able to extract feature maps from non-processed data obtaining high level abstractions over input data by using trained filters [25]. The weights of these filters are corrected during the training process by stochastic gradient descent or any other variation of the gradient descent algorithm.

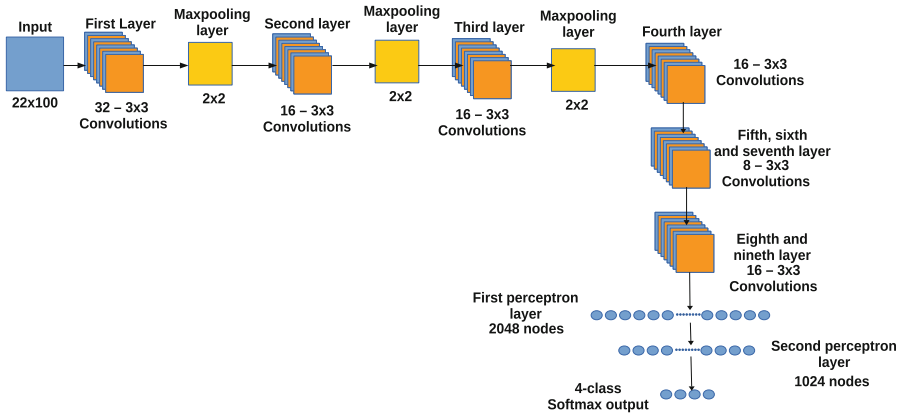


Fig. 4. Configuration of the convolutional neural network.

We found that the lack of convolutional layers can give rise to a low classification performance. As a result of this, the proposed network architecture consists of nine convolutional layers, which are responsible of extracting the necessary features from the preprocessed signals. All convolutional layers are composed by 3×3 kernels. With these kernels 32, 16, 16, 16, 8, 8, 8, 16, 16, features maps are computed on each layer, respectively. The first three layers are followed by MaxPooling sub sampling applied in 2×2 regions to preserve only the most relevant information, Rectified Linear Unit (ReLU) activation was chosen to avoid nonlinearities. Each convolutional layer was initialized by Glorot uniform kernel initializer [26]. Finally, a multilayer perceptron (MLP) was included to receive the information of the last convolutional layer for MI detection. The first dense

layer is composed of 2048 neurons and the last one uses 1024 units. The output of the perceptron is connected to a Dropout function additional layer, in order to avoid over fitting, and finally fed a softmax function layer to classify. The softmax is the output layer of the neural network, which contains a neuron corresponding to each type of MI (class), For a total of 4 neurons. One Dropout layer was added after the last convolutional layer in order to avoid over fitting. Finally, Adaptive Moment Estimation (Adam) optimizer [27] was selected as the optimization algorithm. This method is based on adaptive estimates of lower-order moments due to its computational efficiency and its capacity to work with non-stationary objectives. The parameters of the optimizer follow those provided in Kingma et al. original paper [27], except for the initial learning rate fixed 0.0001.

4 Results and Discussion

4.1 Experimental Setup

Proposed deep learning architecture was implemented in Keras [28], a deep learning framework that uses TensorFlow [29] as backend. A learning model was trained for each subject during 120 epochs, and performance was computed using a 5-fold cross-validation strategy, i.e. 80% of subject samples for training and 20% for testing. Thus, nine evaluations, one per subject, were carried out. Performance measurements are computed as the mean for the five folds per subject, and the mean for all subjects as the model accuracy.

Experiments were carried out in Kubuntu Linux distribution, in a Dell Precision T5810 CPU equipped with Intel Xeon processors of 3 GHz, 8 cores, 64-bit architecture, 16 GB of RAM. No GPU was used in this work, the network was trained only in CPU cores. The preprocessing stage of the database using Matlab was performed on the same computer.

4.2 Experimental Results

The results obtained for each of the PSD estimators are shown in Table 1, which report the mean testing accuracy and Cohen's kappa coefficient per subject, for both Periodogram and Welch PSD estimators, from a 5-fold cross-validation strategy. As can be observed, the Welch function obtained a higher performance compared to the Periodogram function (accuracies of 0.88 and 0.82, respectively), improving in a 13.6% the best accuracy reported in state of the art for the same dataset. It is also important to note the small variability (standard deviation) of the proposed approach in comparison with state of the art methods, as it is shown in Table 2, even to subjects with poor results in previous works such as the subjects 2, 5 and 6.

Additionally, Table 3 presents a comparison of the mean accuracy obtained by the proposed approach using the Welch function and state of the art works, including they that not reported results per subject. According with the results, the proposed approach outperform all of them.

Table 1. Performance of the proposed approach for each subject in the database

Subject	Welch		Periodogram	
	Accuracy	Kappa	Accuracy	Kappa
1	0.86	0.82	0.78	0.78
2	0.87	0.83	0.81	0.75
3	0.93	0.91	0.86	0.82
4	0.85	0.81	0.80	0.73
5	0.84	0.79	0.77	0.69
6	0.86	0.82	0.78	0.71
7	0.88	0.84	0.81	0.75
8	0.89	0.86	0.84	0.78
9	0.92	0.89	0.89	0.86
Mean \pm std	0.88 \pm 0.030	0.84 \pm 0.039	0.82 \pm 0.042	0.75 \pm 0.056

Table 2. Performance comparison with state of the art methods according to accuracy per subject

Subject	Yang et al. (2015)	Sakhavi et al. (2015)	Bashashati et al. (2015)	Periodogram (Proposed)	Welch (Proposed)
1	0.77	0.81	0.79	0.78	0.86
2	0.50	0.54	0.61	0.81	0.87
3	0.80	0.85	0.86	0.86	0.93
4	0.54	0.65	0.74	0.80	0.85
5	0.65	0.59	0.60	0.77	0.84
6	0.49	0.44	0.57	0.78	0.86
7	0.81	0.84	0.87	0.81	0.88
8	0.84	0.87	0.81	0.84	0.89
9	0.82	0.78	0.84	0.89	0.92
Mean Acc. \pm std	0.69 \pm 0.15	0.71 \pm 0.16	0.74 \pm 0.12	0.82 \pm 0.04	0.88 \pm 0.03

Table 3. Performance comparison with state of the art methods according to mean accuracy

Methodology	Mean accuracy
Merinov et al. 2016	0.650
Helal et al. 2017	0.670
Yang et al. 2015	0.692
Sakhavi et al. 2015	0.706
Bashashati et al. 2015	0.743
Proposed approach	0.879

5 Conclusions and Future Works

In this paper a new approach to identify motor imagery from EEG signals was proposed and evaluated. Proposed approach is composed of two main components, a PSD based feature extraction and a deep learning classification. According with the reported results, the proposed method provides a reliable strategy for differentiating the movement intention, outperforming state of the art methods that were evaluated using the same dataset. Additionally, Two PSD estimators were herein evaluated, showing that modified periodogram (Welch function) reach a better representation of the signal variations. We evaluate heuristically other network architectures, changing number of layers and convolutional filters per layer (results not shown). However, we note that architectures composed by more layers but less filters per layer, obtained better performance than those with less layers but many filters.

On the other hand, because EEG signals have not spatial relationship between channels, we implemented a preprocessing step that allows to reordering the location of the EEG channels according to their location in the scalp. Other spatial filters could be explored in this stage to evaluate the advantages and drawbacks of this scheme.

As future work, global optimization algorithms could be employed for hyperparameter optimization of learning model parameters, which could to improve the final performance. In addition, different feature extraction methods could be also evaluated.

References

1. McFarland, D.J., Wolpaw, J.R.: Brain-computer interface operation of robotic and prosthetic devices. *Adv. Comput.* **79**, 169–187 (2008)
2. McFarland, D.J., Wolpaw, J.R.: Brain-computer interfaces for communication and control. *Commun. ACM* **54**(5), 60–66 (2011)
3. Elghrabawy, A., Wahed, M.A.: Prediction of five-class finger flexion using ECoG signals. In: Cairo International Biomedical Engineering Conference (CIBEC). IEEE, pp. 1–5 (2012)
4. Nicolas-Alonso, L.F., Gomez-Gil, J.: Brain computer interfaces, a review. *Sensors* **12**(2), 1211–1279 (2012)
5. Pfurtscheller, G., Neuper, C.: Motor imagery and direct brain-computer communication. *Proc. IEEE* **89**(7), 1123–1134 (2001)
6. Azar, A.T., Balas, V.E., Olariu, T.: Classification of EEG-based brain-computer interfaces. In: Iantovics, B., Kountchev, R. (eds.) *Advanced Intelligent Computational Technologies and Decision Support Systems*. SCI, vol. 486, pp. 97–106. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-00467-9_9
7. Guerra, Z.F., Lucchetti, A.L., Lucchetti, G.: Motor imagery training after stroke: a systematic review and meta-analysis of randomized controlled trials. *J. Neurol. Phys. Ther.* **41**(4), 205–214 (2017)
8. Cong, F., Lin, Q.-H., Kuang, L.-D., Gong, X.-F., Astikainen, P., Ristaniemi, T.: Tensor decomposition of EEG signals: a brief review. *J. Neurosci. Meth.* **248**, 59–69 (2015)

9. Rahman, M., Joadder, M.A.M.: A review on the components of EEG-based motor imagery classification with quantitative comparison. *Appl. Theory Comput. Technol.* **2**(2), 1–15 (2017)
10. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for EEG-based brain-computer interfaces. *J. Neural Eng.* **4**(2), R1 (2007)
11. Shen, Y., Lu, H., Jia, J.: Classification of motor imagery EEG signals with deep learning models. In: Sun, Y., Lu, H., Zhang, L., Yang, J., Huang, H. (eds.) *ISCIIDE 2017*. LNCS, vol. 10559, pp. 181–190. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67777-4_16
12. Brunner, C., Leeb, R., Müller-Putz, G., Schlögl, A., Pfurtscheller, G.: “BCI competition 2008-Graz data set a,” Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology, pp. 136–142 (2008)
13. Bashashati, H., Ward, R.K., Birch, G.E., Bashashati, A.: Comparing different classifiers in sensory motor brain computer interfaces. *PloS One* **10**(6), e0129435 (2015)
14. Helal, M.A., Eldawlatly, S., Taher, M.: Using autoencoders for feature enhancement in motor imagery brain-computer interfaces. In: 2017 13th IASTED International Conference on Biomedical Engineering (BioMed), pp. 89–93. IEEE (2017)
15. Merinov, P., Belyaev, M., Krivov, E.: Filter bank extension for neural network-based motor imagery classification. In: *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE (2016)
16. Sakhavi, S., Guan, C., Yan, S.: Parallel convolutional-linear neural network for motor imagery classification. In: *23rd European Signal Processing Conference (EUSIPCO)*, pp. 2736–2740. IEEE (2015)
17. Ang, K.K., Chin, Z.Y., Wang, C., Guan, C., Zhang, H.: Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Front. Neurosci.* **6**, 39 (2012)
18. Yang, H., Sakhavi, S., Ang, K.K., Guan, C.: On the use of convolutional neural networks and augmented CSP features for multi-class motor imagery of EEG signals classification. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2620–2623. IEEE (2015)
19. Walker, I.: *Deep convolutional neural networks for brain computer interface using motor imagery* (2015)
20. Welch, P.: The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **15**(2), 70–73 (1967)
21. Liavas, A.P., Moustakides, G.V., Henning, G., Psarakis, E.Z., Husar, P.: A periodogram-based method for the detection of steady-state visually evoked potentials. *IEEE Trans. Biomed. Eng.* **45**(2), 242–248 (1998)
22. Jones, E., Oliphant, T., Peterson, P., et al.: *SciPy: open source scientific tools for Python* (2001). <http://www.scipy.org/>
23. Parhi, K.K., Ayinala, M.: Low-complexity Welch power spectral density computation. *IEEE Trans. Circ. Syst. I Regul. Pap.* **61**(1), 172–182 (2014)
24. Cannon, M.J., Percival, D.B., Caccia, D.C., Raymond, G.M., Bassingthwaight, J.B.: Evaluating scaled windowed variance methods for estimating the Hurst coefficient of time series. *Physica A Stat. Mech. Appl.* **241**(3–4), 606–626 (1997)
25. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
26. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256 (2010)

27. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
28. Chollet, F., et al.: Keras (2015). <https://github.com/fchollet/keras>
29. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint [arXiv:1603.04467](https://arxiv.org/abs/1603.04467) (2016)



Long Term Use Effects of a P300-Based Spelling Application

Cristian-Cezar Postelnicu^{1,2}(✉), Florin Gîrbacia^{1,2},
Octavian Machidon^{1,2}, and Gheorghe-Daniel Voinea^{1,2}

¹ Department of Automotive and Transport Engineering,
Transilvania University of Braşov, Braşov, Romania
cristian-cezar.postelnicu@unitbv.ro

² Department of Electronics and Computers,
Transilvania University of Braşov, Braşov, Romania

Abstract. The P300-based Brain Computer Interfaces (BCIs) are widely used for communication purposes and proven to be fast and reliable. The present study aimed to identify the effects of a long term use of a P300-based spelling application for clinical healthy subjects under two display cases represented by a computer monitor and a head-mounted display (HMD). The typical 6×6 row-column spelling application was used by 7 participants under 12 recording sessions for both display methods. The results of the experiment showed that over time there is a statistical significant difference for the average accuracy rates for each display method. The average classification accuracies improved overall by a maximum of 21.43% for monitor display and 25.72% for HMD. However, the results of the experiment demonstrated that no statistical significant interaction existed between the two display methods.

Keywords: P300 speller · Brain computer interface · Electroencephalography

1 Introduction

Brain-computer interfaces (BCIs) allow people to use alternative methods of human-computer interfaces. These types of systems are proven to successfully replace the traditional computer control interfaces represented by keyboard and mouse. They have been initially developed in order to allow patients suffering of severe diseases like locked-in syndrome or amyotrophic lateral sclerosis (ALS) to communicate with others. Nowadays such interfaces can be also used by clinically healthy people as a supplementary communication and control channel.

BCIs are mainly based on the use of electroencephalography (EEG) recordings from the user's scalp, this solution being non-invasive and cost-effective compared with other solutions [1, 2]. Event related potentials (ERPs) extracted from the EEG signals are usually used as key features with various applications for spelling [3–6], drawing [7] or even to control smart houses [8, 9]. The most common ERP is the P300 potential, its usefulness being proven especially for spelling application but not only. The applications based on P300 have proven to work with accuracy levels of even 100% and with high information transfer rates [10].

There are relatively few long-term BCI usage studies, most of them being related to either independent home use of BCI-controlled applications for disabled patients [11–13] or clinical tests involving follow-ups after a certain period of time (usually months) [14, 15]. Also, there are a few studies regarding multiple-day usage of BCI for patients with various disabilities [16].

Apart from the clinical effects on the users, these long-term studies also provide a perspective on the accuracy and amplitude evolution compared to single-time experiments. Hence, while it is possible to obtain high accuracies in single-time experiments (during calibration and training), the results may not be conclusive. Instead, the long-term studies have shown that only after a certain period of time (several sessions, usually after a couple of weeks and months [11]) the accuracy levels tend to reach a certain value and remain stable [17]. It is also true that in the case of users suffering from serious disabilities like ALS, the accuracy levels may decline during a longer period of time (years) due to a decline in the users physical health caused by ALS. However, healthy user groups have shown to maintain a stable accuracy level over the entire long-term period.

Regarding the amplitude, studies have shown it to vary during longer period of times, with a short increase followed by a prolonged decrease in the study presented in [17].

Overall, the reviewed relevant research supports the fact that the long-term use of BCIs leads to high levels of accuracy and successful training for all healthy users and the vast majority of patients also (depending on their level of disablement due to disease). Also, studies have underlined the importance of ease of use, reliability and easy adjustment of the BCI system in the case of long-term usage [11, 12].

An important issue revealed by [16] is related to the need for spatially and frequency-range stable neural signs used for control, since it is not feasible to relearn the control parameters before every BCI session. Hence, it is important in the case of long-term applications to have as control features signals that are robust and stable over long periods of time [16].

Regarding potential side-effects of long-term usage, no significant reactions were found, just minor headaches or discomfort caused by fatigue in few cases [14].

One important BCI application is in the field of embodiment, i.e. using BCIs to control a robot (a humanoid robot or a robotic body part). Work in this direction shows that robotic humanoid whole or part body control can be achieved by using an (EEG) – based BCI [18]. However, one of the challenges encountered in this type of applications is related to reduced robot steering accuracy leading to the impossibility of positioning the robot in a precise position in space [19]. The authors of [19] propose a solution to improve BCI embodiment applications by using a Head-Mounted Display (HMD) to provide assisted steering, navigation and interaction through augmented reality.

Virtual reality (VR) is an emerging domain with multiple applications in almost all possible research fields. Also, it is currently completed by the mass production of commercial head-mounted displays (HMDs) like Oculus Rift [20], HTC VIVE [21] and Samsung Gear VR [22]. Virtual reality has already been used in BCIs and accurate results have been recorded. It was successfully used for games [23], virtual worlds navigation [24] and even for spelling applications [10].

It has been shown that applications fostering the synergy between HMD technology and BCIs have the potential to improve user experience and enable new types of immersive applications [25]. This is because such a symbiosis promises to surpass the common disadvantages of both technologies: HMDs can be optimized using the EEG signals from the BCI (e.g. for achieving 100% accuracy in detecting head rotation, like in [25]) and BCI systems can use the HMD as an adaptive screen to improve user experience during training on the fly.

BCI and HMDs are already used in applications shown to improve the control of 3D objects in VR environments. In [26] the authors proposed a system integrating an Oculus Rift HMD, an eye tracking system and a BCI interface that allows users to point the objects they are interested in by eye gazing and control them by thinking.

Moreover, BCIs and HMDs can be used to determine the quality of a VR application or scenario by monitoring the brain signal activity and comparing the results when the user observes real world and VR objects [27].

The current study aims to identify the effects of long term use of P300-based spelling applications for clinical healthy subjects in two display cases represented by a computer monitor and a HMD. The study is part of a research project which aims to offer feasible BCI-based communication and control channels for clinical healthy subjects for long-term use. Most studies are usually focused on single short use of spelling applications, while very few investigated the effects for a long-term purpose.

2 Materials and Method

2.1 Participants

Seven healthy, subjects (6 male and 1 female, mean age = 27.2 years, range = 22–40 years) with no prior experience in EEG or BCI were recruited as participants, while a single participant previously experimented HMDs. None of the subjects reported a history of psychiatric or neurologic symptoms. All the participants gave their informed consent and none of them was remunerated for the participation in the study.

2.2 Experimental Design and Procedure

The typical spelling matrix of 6×6 rows and columns filled with the 26 English letters, digits from 1 to 9 and the space character was used for the experiment [3]. The chosen flashing method was the row-column paradigm (RC), high accuracy classification rates being achieved in this case [4]. The RC paradigm was set to highlight items for 100 ms (flash time) with a short time between flashes of 60 ms (dark time) while all characters are grey. The number of repetitions was set to 7. This yields an interval of 13.44 s ($6 \text{ rows} \times 160 \text{ ms} \times 7 \text{ flashes} + 6 \text{ columns} \times 160 \text{ ms} \times 7 \text{ flashes}$) to select a character. After each character selection the matrix stopped flashing for 4 s (spelling pause - SP) which is enough time for the user to visualize the last selected item and to focus on the next character to be spelled.

The subjects were instructed to look at the character prompted to be spelled and to silently count each time it was highlighted. EEG signal patterns are known to be

influenced by factors like fatigue, motivation, mental state, level of attention or motivation [1]. Thus, before each experiment the users were requested to perform a calibration session to ensure that the classification was performed based on the users' current EEG recorded data.

Each subject was asked to spell a 4-character text, one letter at a time, without feedback from the P300 spelling application. The calibration data were then processed by linear discriminant analysis (LDA) to determine coefficients for online classification. Next, the users were asked to spell a 10-character. After each character selection the LDA was applied on the EEG data for each column and row. The application provided feedback to the user by indicating the character identified by the classifier (copy-spelling presentation method). The 10-character sequence used for copy-spelling was always different for each participant and balanced to cover as much as possible all the symbols in the 6×6 matrix.

The study was spanned over 12 weeks of recordings aiming to identify the effects of long-term use of a P300 spelling application. Two display cases have been considered for this study, the former represented by a 24 inches computer monitor (60 Hz), while the later by an Oculus Rift HMD system (see Fig. 1 for system architecture). The subjects were asked to participate for the long-term analysis for both display cases during each week, allowing them between 5 to 8 days between two consecutive recording sessions. A recording session consisted of a calibration session and two separate tests, one for the monitor display ("monitor") case and one for the VR system ("Oculus"), performed in a counter balanced manner to ensure there are few correlations between them.

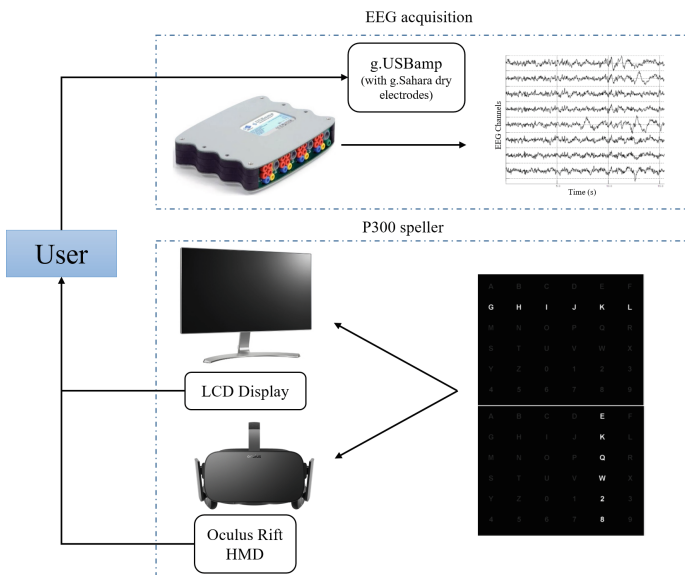


Fig. 1. System architecture

2.3 Data Acquisition, Calibration and Classification

The EEG signals were measured by eight g.Sahara active dry electrodes positioned at Fz, Cz, P3, P4, POz, PO7, PO8 and Oz according to the 10/20 International System. All channels were referenced to the right earlobe and the ground electrode was positioned at FPz. The data were recorded by a g.USBamp (g.tec Medical Engineering GmbH, Austria) biopotentials amplifier and sampled at 256 Hz. Also, the data were filtered using a high pass of 0.5 Hz and a low pass of 30 Hz. A notch filter around 50 Hz was used to reject the power line noise.

For each session the subjects were requested to perform the calibration procedure lasting maximum 3 min during which they were spelling a 4-character word. The data acquired during the calibration session is down-sampled to 64 Hz and the EEG signals are segmented in 800 ms epochs starting at each highlighted stimulus. The extracted trial epochs for all stimuli are entered in the LDA classifier to derive the weighting coefficients, separating them into two classes: target and non-target. During the online copy-spelling task the LDA will associate each visual stimulus with one of the two classes. Finally, the LDA will choose the character with the highest sum of the weighted parameters.

2.4 Effectiveness and Efficiency Metrics

One of the most used and relevant efficiency metric is the online spelling accuracy. It is calculated by dividing the number of correctly spelled characters by all characters that needed to be spelled. For the present study online classification accuracy rates have been recorded for both display methods, “monitor” and “Oculus”, for each recording session and each participant. A statistical analysis was performed to identify whether significant correlations between the display methods and time were present.

Another important measure that is used to grade the efficiency of the system is the Information Transfer Rate (ITR). Firstly, Shannon’s [28] formula was applied to calculate the bitrate:

$$B = \log_2 N + P \log_2 P + (1 - P) \log_2 \left(\frac{1 - P}{N - 1} \right) \quad (1)$$

where: N is the number of possible selections (36 in case of the 6×6 matrix) and P represents the probability that the desired item will actually be selected (classification accuracy). The bitrate is then multiplied by the number of possible decisions per minute (M) to obtain the ITR in bits/min.

Parameter M is calculated from:

$$M = 60 / (N_{rep} \times N_{groups} \times ISI + N_{groups} \times SP) \quad (2)$$

where: N_{rep} represents the number of repetitions for each character (7 repetitions), N_{groups} is the number of groups (6 rows and 6 columns), ISI stands for the time interval between two consecutive flashes (100 ms flash time + 60 ms dark time) and SP represents the spelling pause between two consecutive characters (36 s = 9 pauses \times 4 s).

The total task completion time was calculated as the total number of flashes used for the copy-spelling task, also including the pauses between the selections of two consecutive characters ($170.4\text{ s} = 13.44\text{ s} \times 10\text{ characters} + 36\text{ s}$).

3 Results

Average accuracies obtained for the spelling tasks with the two display modalities over the entire period are grouped per session and are presented in Fig. 2 and Table 1. The average accuracy rate for the entire period was of 86.07% for “monitor” and 83.21% for “Oculus” display methods. The ITR had an average value of 16.54 bits/min (SD = 1.47) for “monitor” and 16.23 bits/min (SD = 1.54) for “Oculus”, with a minimum of 13.19 bits/min and a maximum of 18.2 bits/min for both of them.

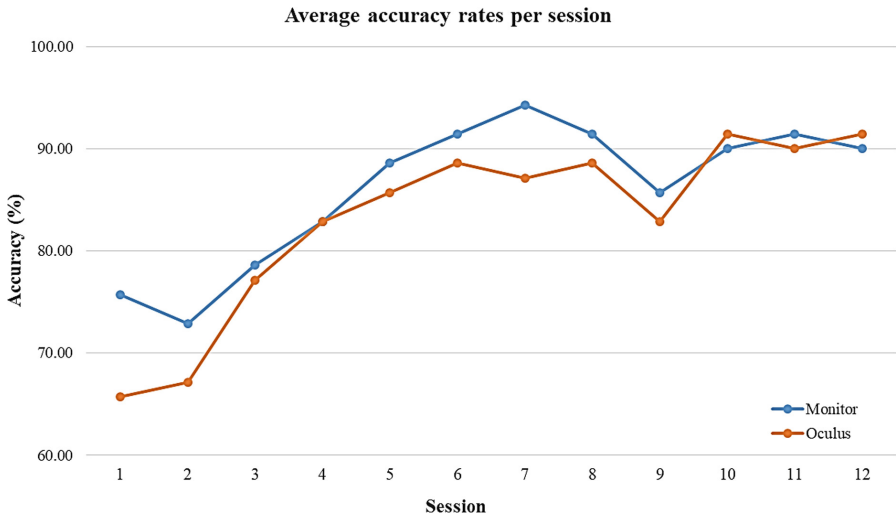


Fig. 2. Average accuracies for both display methods and all recording sessions

The repeated measures ANOVA revealed a main effect over “time” $F(4.821, 57.857) = 7.138$, Greenhouse-Geisser corrected $p < 0.05$, but no main effect of “display” on spelling performance $F(1, 12) = 0.509$, $p = 0.489$ and no significant interaction between “time” and “display” $F(4.821, 57.857) = 0.338$, $p = 0.882$.

For the “monitor” display case statistically significant differences have been found during “time” especially between sessions $S1 \div 2$ compared with $S5 \div 8$, $S10 \div 12$ ($p < 0.05$), $S3$ compared with $S5 \div 8$, $10 \div 11$ ($p < 0.05$), and $S4$ and $S6$ ($p < 0.05$) (see Table 1 for accuracy values). For the “Oculus” case statistically significant differences have been found between sessions $S1$ and $S5 \div 8$, $10 \div 12$ ($p < 0.05$), $S2$ and $S4 \div 8$, $10 \div 12$ ($p < 0.05$), and $S3$ and $S6$, $10 \div 12$ ($p < 0.05$).

Figure 3 displays the ERP waveforms for one participant for all electrodes under “monitor” condition for targets and non-targets to facilitate comparison. By analyzing

Table 1. Average accuracies for the two display methods and all recording sessions

Session	Display method			
	Monitor		Oculus	
	Average accuracy (%)	Standard deviation	Average accuracy (%)	Standard deviation
S1	75.71	15.91	65.71	14.00
S2	72.86	17.50	67.14	8.81
S3	78.57	14.57	77.14	14.85
S4	82.86	13.85	82.86	13.85
S5	88.57	13.55	85.71	10.50
S6	91.43	9.90	88.57	12.45
S7	94.29	4.95	87.14	10.30
S8	91.43	6.39	88.57	9.90
S9	85.71	14.00	82.86	16.66
S10	90.00	9.26	91.43	11.25
S11	91.43	6.39	90.00	5.35
S12	90.00	9.26	91.43	9.90

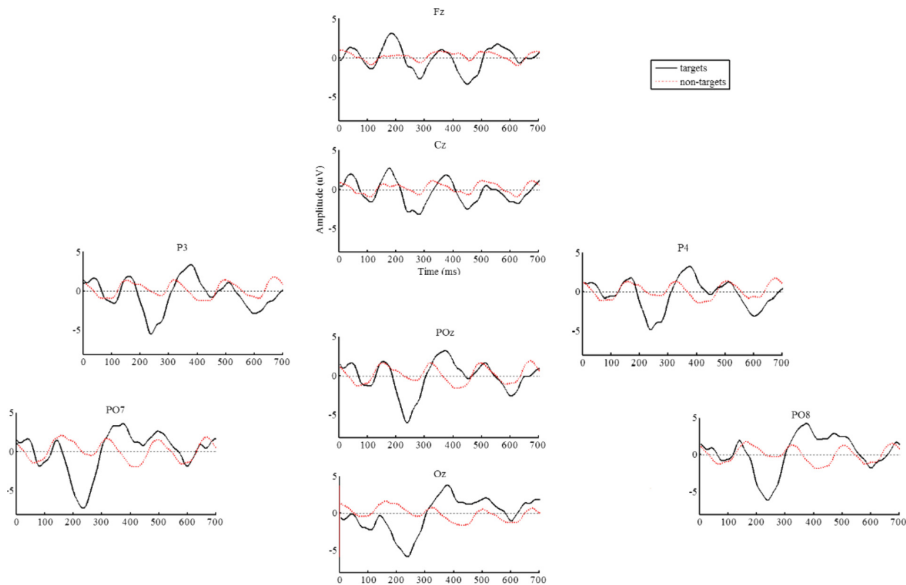


Fig. 3. ERP waveforms average for targets and non-targets for one user under “monitor” case

all the waveforms obtained for all the participants it was found that the P300 potential was most visible in the time frame between 250 and 500 ms after the stimulus onset. Also, the P300 had higher amplitude values for the P3, P4, POz and Oz channels in general, but for most of the participants the POz channel yielded the highest values being further analyzed for amplitude and latency parameters.

The ANOVA yielded no significant results for the P300 latency on the POz channel for “time” $F(11, 55.222) = 1.439, p = 228$, Greenhouse-Geiser corrected, nor for “display” $F(1, 12) = 0.038, p = 0.849$, nor a significant interaction between “time” and “display” $F(4.602, 55.222) = 1.321, p = 0.271$. The P300 latency had an average value of 343.65 ms ($SD = 37.85$) for “monitor” and 357.59 ms ($SD = 36.17$) for “Oculus” display methods.

The P300 amplitude on the POz channel had an average value of 3.53 μV ($SD = 1.33$) for “monitor” and 3.65 μV ($SD = 1.23$) for “Oculus” display methods. On average the “Oculus” display method was higher by 0.12 μV compared with the “monitor” method, but no significant effects have been found during “time” $F(11, 132) = 1.084, p = 0.379$ nor for the “display” method $F(1, 12) = 0.025, p = 0.878$.

4 Discussion

The study intended to discover whether significant effects can be found after a long term use of a P300 speller application under two display methods conditions. It was shown that the average accuracies achieved during online spelling did not statistically differ between the Oculus Rift VR and the 22 inches monitor even if the experiment was spanned over 12 recording sessions.

By analyzing the obtained results over the 12 sessions it can be seen that for both display cases the average accuracy rates increased between sessions achieving a maximum average of 94.29% for “monitor” case during S7 and 91.43% for “Oculus” case during S10 and S12 (see Table 1).

In general, most of the participants were comfortable using both Oculus Rift and EEG cap. One of the participants didn’t feel too comfortable during S1 ÷ 3 with Oculus Rift and/or the electrodes, but in time he got used with them. He managed to achieve an average accuracy of 71.66% ($SD = 16.24$), with a minimum of 50% during S1 and a maximum of 90% during S7. Also, it should be stated that for a few sessions a couple of participants took a while to accommodate with the g.Sahara electrodes.

For S9 the average accuracy decreased by 5.72% for “monitor” and 5.71% for “Oculus” cases, due to the poor performance of two participants which declared to have problems to concentrate on the flashing items.

Two of the participants managed to achieve similar or even higher accuracies during the “Oculus” test for at least 8 sessions compared with the “monitor” case, even if only one of them had previous experience with HMDs.

All the participants managed to achieve during the entire experiment a maximum accuracy of at least 90%, while six of them reached 100% for 2 up to 5 sessions for both display methods.

The P300 amplitude and latency yielded on the POz channel the highest discriminative ERP values for all the participants, but no statistical significant effects were identified between the two display methods or sessions.

5 Conclusions

With the current study we were able to identify a few correlations that arise between two display methods for a long term experiment with a P300-based speller application. No statistical significant differences have been found between the “monitor” and “Oculus” display methods over time. Separately, each display method presented significant effects on the accuracy rates mainly by comparing initial ($S1 \div 3$) with middle ($S5 \div 8$) and last ($S10 \div 12$) recording sessions. This demonstrates that on average participants managed to increase the spelling accuracy during time. It was also demonstrated that most of the participants, 6 out of 7, can achieve the maximum spelling accuracy (100%) for both HMD and monitor display methods.

Acknowledgments. This work was supported by a grant of the Ministry of National Education and Scientific Research, RDI Programme for Space Technology and Advanced Research - STAR, project number 566.

References

1. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. *Clin. Neurophysiol.* **113**(6), 767–791 (2002)
2. Allison, B.Z., Leeb, R., Brunner, C., Muller-Putz, G.R., Bauernfeind, G., Kelly, J.W., Neuper, C.: Toward smarter BCIs: extending BCIs through hybridization and intelligent control. *J. Neural Eng.* **9**(1), 7 (2012)
3. Donchin, E., Spencer, K.M., Wijesinghe, R.: The mental prosthesis: assessing the speed of a P300-based brain-computer interface. *IEEE Trans. Rehabil. Eng.* **8**(2), 174–179 (2000)
4. Guger, C., Daban, S., Sellers, E., Holzner, C., Krausz, G., Carabalona, R., Gramatica, F., Edlinger, G.: How many people are able to control a P300-based brain-computer interface (BCI)? *Neurosci. Lett.* **462**(1), 94–98 (2009)
5. Postelnicu, C.-C., Talaba, D.: P300-based brain-neuronal computer interaction for spelling applications. *IEEE Trans. Biomed. Eng.* **60**(2), 534–543 (2012)
6. Townsend, G.T., LaPallo, B.K., Boulay, C.B., Krusienski, D.J., Frye, G.E., Hauser, C.K., Schwartz, N.E., Vaughan, T.M., Wolpaw, J.R., Sellers, E.W.: A novel P300-based brain-computer interface stimulus presentation paradigm: moving beyond rows and columns. *Clin. Neurophysiol.* **121**(7), 1109–1120 (2010)
7. Botrel, L., Holz, E.M., Kubler, A.: Brain painting V2: evaluation of P300-based brain computer interface for creative expression by an end-user following the user-centered design. *Brain Comput. Interfaces* **2**, 135–149 (2015)
8. Edlinger, G., Holzner, C., Guger, C.: A hybrid brain-computer interface for smart home control. In: Jacko, J.A. (ed.) *HCI 2011. LNCS*, vol. 6762, pp. 417–426. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21605-3_46
9. Postelnicu, C.-C., Covaci, A., Panfir, A.N., Talaba, D.: Evaluation of a P300-based interface for smart home control. In: Camarinha-Matos, L.M., Shahamatnia, E., Nunes, G. (eds.) *DoCEIS 2012. IAICT*, vol. 372, pp. 179–186. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28255-3_20
10. Kathner, I., Kubler, A., Halder, S.: Rapid P300 brain-computer interface communication with a head-mounted display. *Front. Neurosci.* **9**, 207 (2015)

11. Holz, E.M., Botrel, L., Kaufmann, T., Kübler, A.: Long-term independent brain-computer interface home use improves quality of life of a patient in the locked-in state: a case study. *Arch. Phys. Med. Rehabil.* **96**(3), S16–S26 (2015)
12. Sellers, E.W., Vaughan, T.M., Wolpaw, J.R.: A brain-computer interface for long-term independent home use. *Amyotrophic Lateral Scler.* **11**(5), 449–455 (2010)
13. Shahriari, Y., Vaughan, T.M., Corda, D.E., Zeitlin, D., Wolpaw, J.R., Krusienski, D.J.: EEG correlates of performance during long-term use of a P300 BCI by individuals with amyotrophic lateral sclerosis. In: *Proceedings of the Fifth International Brain -Computer Interface Meeting* (2013). <https://doi.org/10.3217/978-3-85125-260-6-29>
14. Onose, G., Grozea, C., Anghelescu, A., Daia, C., Sinescu, C.J., Ciurea, A.V., Popescu, C.: On the feasibility of using motor imagery EEG-based brain-computer interface in chronic tetraplegics for assistive robotic arm control: a clinical test and long-term post-trial follow-up. *Spinal Cord* **50**(8), 599 (2012)
15. Nandrajog, P., Idris, Z., Azlen, W.N., Liyana, A., Abdullah, J.M.: The use of event-related potential (P300) and neuropsychological testing to evaluate cognitive impairment in mild traumatic brain injury patients. *Asian J. Neurosurg.* **12**(3), 447 (2017)
16. Blakely, T., Miller, K.J., Zanos, S.P., Rao, R.P., Ojemann, J.G.: Robust, long-term control of an electrocorticographic brain-computer interface with fixed parameters. *Neurosurg. Focus* **27**(1), E13 (2009)
17. Botrel, L., Holz, E.M., Kübler, A.: Using brain painting at home for 5 Years: stability of the P300 during prolonged BCI usage by two end-users with ALS. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) *AC 2017. LNCS (LNAI)*, vol. 10285, pp. 282–292. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_20
18. Gergondet, P., Kheddar, A., Hintermüller, C., Guger, C., Slater, M.: Multitask humanoid control with a brain-computer interface: user experiment with HRP-2. In: Desai, J., Dudek, G., Khatib, O., Kumar, V. (eds.) *Experimental Robotics. Springer Tracts in Advanced Robotics*, vol 88. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-319-00065-7_16
19. Petit, D., Gergondet, P., Cherubini, A., Meilland, M., Comport, A.I., Kheddar, A.: Navigation assistance for a BCI-controlled humanoid robot. In: *2014 IEEE 4th Annual International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 246–251. IEEE (2014)
20. Oculus Rift. www.oculus.com/rift/. Accessed 09 Jan 2018
21. HTC VIVE. www.vive.com. Accessed 09 Jan 2018
22. Samsung Gear VR. www.samsung.com/global/galaxy/gear-vr/. Accessed 09 Jan 2018
23. Finke, A., Lenhardt, A., Ritter, H.: The MindGame: a P300-based brain-computer interface game. *Neural Netw.* **22**(9), 1329–1333 (2009)
24. Groenegrass, C., Holzner, C., Guger, C., Slater, M.: Effects of P300-based BCI use on reported presence in a virtual environment. *Presence* **19**(1), 1–11 (2010)
25. Brouwer, A.M., van der Waa, J.S., Hogervorst, M.A., Cacace, A., Stokking, H.: A feasible BCI in real life: using predicted head rotation to improve HMD imaging. In: *Proceedings of the 2017 ACM Workshop on an Application-Oriented Approach to BCI Out of the laboratory*, pp. 35–38. ACM (2017)
26. Chun, J., Bae, B., Jo, S.: BCI based hybrid interface for 3D object control in virtual reality. In: *2016 4th International Winter Conference on Brain-Computer Interface (BCI)*, pp. 1–4. IEEE (2016)
27. Chun, J., Kaongoen, N., Jo, S.: EEG signal analysis for measuring the quality of virtual reality. In: *15th International Conference on Control, Automation and Systems (ICCAS)*, pp. 1801–1804. IEEE (2015)
28. Shannon, C.E., Weaver, W.: *The Mathematical Theory of Communication*. University of Illinois Press, Champagne, IL (1964)



A Wearable Multisensory, Multiagent Approach for Detection and Mitigation of Acute Cognitive Strain

Phase I - Vocalization analysis

Anil Raj¹(✉), Brooke Roberts², Kristy Hollingshead¹,
Neil McDonald², Melissa Poquette², and Walid Soussou²

¹ Florida Institute for Human and Machine Cognition,
40 South Alcaniz Street, Pensacola, FL 32503, USA
ara.j@ihmc.us

² Quantum Applied Science and Research,
5754 Pacific Center Blvd., Suite 203b, San Diego, CA 92121, USA

Abstract. While operators performing tasks with high workload can increase task performance in response to limited increases in cognitive stress, chronic or rapidly accelerating stress can exceed the operator's ability to compensate, generating acute cognitive strain (ACS). ACS represents a state wherein performance, situation awareness and cooperativity deteriorate markedly, leading to critical errors, mishaps or casualties. Nearly two decades of augmented cognition (AugCog) research has demonstrated the utility of psychophysiological sensing and analysis for identification and tracking of changes in cognitive state and to modulate human machine interactions for improving system task performance. The proposed approach leveraged prior efforts to modulate cognitive stress using a multiagent approach to acquire and analyze multiple Psychophysiological sensory channels, including changes in vocalizations, to create a reliable and non-intrusive Detector of Acute Cognitive Strain (DACS). The DACS system provides an integrated wearable multi-modal Research Sensor Suite (RSS) using the open-source Adaptive Multiagent Integration (AMI) architecture, that includes analysis agents for electroencephalograph (EEG), electromyography (EMG), video oculography (VOG), vocalization, and others to identify and correlate physiological signatures with cognitive stress and strain. An online AMI agent-based processing algorithm was developed and applied to audio communications to evaluate for changes in speaker vocalization fundamental frequency (F0) and cadence (utterances per minute). This paper describes initial phase results of aerospace mishap vocalization stress marker detection, a potential element of the proposed DACS system. DACS could use these markers to trigger adaptive automation agents that reduce task load and allow pilots to prevent or recover from ACS episodes.

Keywords: Acute cognitive strain · Augmented cognition
Multiagent systems · Vocalization · Stress markers

1 Introduction

Operators under high workload demands can experience cognitive stress and, in the increased tempo and round-the-clock nature of military operations, this stress can lead to “Acute Cognitive Strain” (ACS), wherein performance, situation awareness and cooperativity yield, resulting in critical errors, loss of equipment, or casualties. It is critical, therefore, to develop a warfighter-centric approach to reliably detect ACS in order to mitigate its effects, either through training, or through situation-specific compensatory mechanisms. In order to construct a Detector of Acute Cognitive Strain (DACS), the research team developed a prototype design that combined a wearable multi-modal Research Sensor Suite (RSS) with the Adaptive Multiagent Integration (AMI) software architecture. This combination allows for a DACS system that can be tailored to meet the requirements of specific operational domains. The AMI implementation manages data flow between the various psychophysiological and environmental sensors and the machine learning algorithms to provide real-time estimates of cognitive state in environmental context. Although no reliable methods currently exist for non-intrusive monitoring for ACS, Quantum Applied Science and Research (QUASAR) and the Florida Institute for Human and Machine Cognition (IHMC) addressed the major DACS development risks as part of this Phase I project by identifying and correlating detectable physiological signatures with markers of cognitive strain. While this effort integrated a full suite of proposed sensors for initial RSS test and evaluation, the next logical step would include a down-selection to the minimal set of wearable sensor modalities needed for reliable and robust DACS system. Along with the sensor integration effort, the team investigated and developed an ACS inducing aviation simulation. The design process consisted of parallel development of the multiagent integrated RSS and evaluation of the constituent components by analyzing previously acquired de-identified or publicly available data from high stress mishaps for characteristics of ACS.

1.1 Psychophysiological Research Sensor Suite (RSS) Integration

Building on prior experience and development of augmented cognition (AugCog) systems [1], the authors first reviewed both traditional and novel psychophysiological sensor technologies for Research Sensor Suite (RSS) utility and feasibility of potential for aerospace environment operational use. By using a combination of measures, the classification accuracy for detection of ACS using QUASAR algorithms (QStates; [2]) could potentially improve over single sensor modality detection. The review process identified the following potential sensors for use in either the RSS, the DACS system or both.

Electroencephalography (EEG). Several aspects of EEG can provide measures of mental workload or cognitive stress. Evoked response potential (ERP) components, including differences in P300 latency [3–5], N200 [3, 5], and feedback-related negativity (FRN) amplitudes [6] manifest with acute cognitive stress. Neural responses to errors can indicate the quality of stress regulation. For instance, clear detection of Stroop task errors observed in FRN predicts less cortisol during task, indicating

successful stress regulation; while greater error-related alpha suppression (ERAS) predicts increased cortisol levels during task and may reflect error-related arousal and stress that is not adaptive [7]. While the effects of challenging cognitive tasks on ERPs are well-known, the temporal precision and data quality required for reliable detection of stressful mental states based upon these ERPs have proven less operationally feasible than other spectral EEG features (e.g., changes in power across multiple frequency bands), which require less temporal precision to time-locked events. Psychosocially stressful computer game playing can induce increases in theta, lower alpha, and gamma power, and decreases in upper alpha and beta power [8] that correspond to increases in heart rate and galvanic skin response. Combining these non-specific features with power ratios, frontal asymmetries, and cross-frequency coupling may provide additional support for identifying ACS. Frontal alpha asymmetry increases, with relatively greater left frontal activity, during a stressful working memory task with threat of electroshock [9] and midline frontal theta/midline parietal theta ratios increase along with heart rate (HR) during stressful arithmetic tasks [10]. In addition, the ratio of low frequency power/high frequency power increases along with low frequency power and HR, while high frequency power decreases during stressful video games [11]. Similar EEG features, including alpha asymmetry and alpha/beta ratios have been incorporated into cognitive stress detection systems [12], and frontal delta-beta coupling decreases during Stroop interference (threatening words) emotional stress [13].

Functional Near Infrared (fNIR). fNIR provides measures of brain activity based upon changes in blood hemoglobin oxygenation (HbO₂). fNIR has been incorporated into studies of job-related stress [14], showing that changes in frontotemporal cortex HbO₂ correlate to elevated job stress. Stressful cognitive tasks, including mental arithmetic, are associated with increased bilateral activation of frontal cortex [15–18], selective increases in HbO₂ in left frontal lobe [16], or asymmetrical activations of left and right frontal cortex [19, 20].

Electromyography (EMG). Electrical activity associated with muscle activation can arise from volitional movements and involuntary responses. For the RSS and DACS systems, surface electrodes can acquire these signals reliably and unobtrusively. For aerospace applications, both postural and laryngeal EMG hold promise.

Postural Electromyography (EMG). Stressful cognitive tasks can induce increased upper left and right trapezius muscle activity without affecting the neck musculature [21–26]. Other stressors, including stressful work environments [27, 28] and anticipation of electroshock delivery [29] have reported similar increases in upper trapezius EMG. Thus, tracking the relationship between neck and trapezius EMG could provide a robust measure of a range of cognitive, environmental, and physical stressors.

Laryngeal EMG (LEMG). Certain types of stress, including social stress [30] and elevated autonomic nervous system (ANS) activity [31] can demonstrate increased infrahyoid (IH) surface EMG during public speaking and correlates with increased fear in stress reactivity paradigms in introverted individuals [32]. Surface EMG activity of trapezius and subcutaneous laryngeal muscles also can increase with elevated ANS activity during a cold pressor task (submerging the participant's hand in ice water [31]).

While current research has not validated specific LEMG features as markers of cognitive strain, the above studies indicate that increased surface LEMG activity, particularly the IH region, is generally associated with increases stress.

Vocalization and Speech. A number of acoustic features of speech have shown usefulness for the detection speech under stress, including fundamental frequency (F0, or pitch), formant frequencies (a concentration of acoustic energy around a particular frequency in the speech signal), spectral composition, particularly spectral tilt (which estimates the difference in energy between the 1st and 2nd formants [33, 34]), the maximum and mean intensity of an utterance (measured in decibels [35]), and features based on Mel-Frequency Cepstral Coefficients (MFCCs) [36–39]. Vocalization F0 (i.e., the rate in Hz that the vocal chords open and close) correlates with levels of stress, making it useful beyond simple binary stressed/unstressed classification [40]. Changes in fundamental frequency can indicate human stress or workload level, with higher workload levels being associated with increases in fundamental frequency [41]. Typical fundamental frequency for males is 85–180 Hz and for females it is 165–255 Hz [42]. The United States National Transportation Safety Board [43] has used the following guidelines to approximate pilot stress level in aviation accident investigations:

- An increase in fundamental frequency by 30% (compared with that individual’s speech in a relaxed condition) is characteristic of “stage 1” level of stress, which could result in increased focused attention and improved performance;
- An increase in fundamental frequency by 50–150% is characteristic of “stage 2” level of stress, which could result in the speaker’s performance being hasty and abbreviated and thus degraded; however, the speaker’s performance would not likely display gross mistakes;
- An increase in fundamental frequency by 100–200% is characteristic of “stage 3” level of stress, or panic, which would likely result in the speaker’s inability to think or function logically or productively.

In addition to acoustic measures of the speech signal, temporal aspects of speech, such as number of pauses and total pause duration [44], can provide insight into a subject’s mental state, such as discriminating between cognitively intact individuals and those with cognitive impairment [45], and thus may also prove useful for assessing ACS. While the speech signal clearly carries information on the stress of a speaker, the noisy operational aviation environment itself can reduce the signal-to-noise (SNR) ratio by directly impacting speech (e.g., the Lombard effect [46, 47]). To maximize the SNR, the RSS uses a throat mounted microphone (throat mic) to collect vocalizations and boost reliability for fusion with the other DACS system measures.

Oculography. A number of cognitive state markers manifest with changes in voluntary and involuntary activity of the eyes. Consequently, different technologies have been developed to track eye movements, blinks and pupillary responses.

Electro-oculography (EOG) and Video-oculography (VOG). EOG/VOG can provide some insight into cognitive stress through analysis of blink rate, blink interval (time between two successive eye blinks), and blink duration. Blink patterns differ for fatigue and workload. For instance, eye blink rate increases with drowsiness [48] or hypoxia [49].

Increased workload from increased visual stimuli processing has been associated with increased blink interval and decreased blink duration [50]. Increased blink rates manifest in combined driving and auditory tasks [51], while decreased blink frequency and shorter blink duration have been reported in surgeons performing stressful surgeries [52]. These results indicate that blink patterns differ based upon the nature of the cognitive stressor. Eye movements can correspond with increased heart rate [54] in stressful tasks to reveal threat bias (e.g., sustained attention and delayed disengagement [53]).

Pupillometry. Pupil diameter can also be used as a simple measure of increased workload and ACS, with increased pupil diameter associated with limited availability of cognitive resources. For example, increasing workload in an N-back memory task [55] increases pupil diameter, along with other measures of ACS, including increased heart rate, decreased accuracy, increased frontal cortex activity, and decreased HRV [56]. Likewise, acute stress induced by the Trier Social Stress Test (TSST) task [57] or distressing movies and pictures [58] increases pupil diameter immediately following stress induction. Pupil diameter has also been used in the detection of mental stress vs. relaxation during computer work, with classification accuracy at an average of 85%, compared to 60% for classification based upon GSR results [59]. Taken together, these studies suggest that eye tracking and pupillometry, in particular, could act as simple, useful and reliable sensors for the detection of ACS.

Heart Rate Variability (HRV) and Heart Rate (HR). Increased HR has been commonly reported across multiple stressors, including mental arithmetic [60] and increased workload [61]. Changes in HRV have also provided reliable measures for cognitive stress, with decreased HRV during mental arithmetic [60, 62], Stroop task [62], or in response to increased mental workload [61]. HRV has also been used to classify acute stress during mental tasks compared to baseline or neutral tasks with 80% accuracy [63]. These and other studies have established the reliability of changes in HR and HRV as measures of cognitive stressors.

Pulse Oximetry (PulseOx). Measures of oxyhemoglobin concentration in the blood and measures of “pulse waves” have proposed links to cognitive stress or mental workload. While pulse oximetry is a relatively inexpensive and easy to use method, the literature associating pulse oximetry measures with cognitive strain remains sparse. Minakuchi et al. [64] reported increased pulse rate and decreased finger pulse wave amplitude during a Stroop task, whereas Ahlund et al. [65] reported an increase in pulse wave amplitude for mental arithmetic and cold pressor tasks. While pulse oximetry may be a relatively simple modality to implement, its potential role would likely fall in the realm of detection of ACS in physiologic events (PEs) due to hypoxic stress.

Electrodermal Response (EDR) or Galvanic Skin Response (GSR). Electrodermal response (EDR) or Galvanic skin response (GSR) has commonly been used to measure several aspects of stress. For example, the TSST has induced significant increases in GSR [66, 67]. Similarly, the Mannheim Multicomponent Stress Test (MMST), which simultaneously combines cognitive, emotional, acoustic, and motivational stressors, has induced significant increases in electrodermal activity corresponding to increased cortisol levels, subjective stress ratings, and heart rate. Electrodermal inputs have also been used in a classifier aimed to detect physiological stress with 90% accuracy [67].

While GSR is relatively easy and cost-effective to implement, it lacks specificity and perhaps sensitivity with respect to subtle changes in ACS. In addition, the typical placement of sensors on the volar surface of the fingers can interfere with many operational (e.g., piloting) tasks that require unencumbered hand and finger movement.

2 Prototype RSS and DACS Development

QUASAR and IHMC integrated two prototype hardware RSS variants using AMI for software integration and data synchronization. The first provides EEG (21 channels) plus GSR, ECG and respiration rate using a modified QUASAR dry electrode DSI-24 system, dubbed “ExG” (see Fig. 1). The second variant replaces the ExG with an eight channel QUASAR dry electrode fNIR/EEG headset. Both systems also integrated QUASAR dry electrode ECG and EMG (laryngeal and trapezius) and a commercial off the shelf (CoTS) GSR sensor (Bioderm 2701, UFI, Morro Bay, CA). The RSS integrates the QUASAR sensors with CoTS sensors for pulse oximetry (Radical-7, Masimo Corp., Irvine, CA), eye tracking/pupilometry using a machine vision camera (A33, JeVois, Inc., Los Angeles, CA) running a modified open source eye tracking algorithm [68], and a Peltor tactical throat microphone (MT96-01, 3 M Co., Maplewood, MN).



Fig. 1. RSS prototype sensor hardware and software. A: (top row, left to right) Trapezius muscle EMG and chest strap of ECG/respiration sensor; neck electrodes and throat microphone; GSR electrodes and pulse oximeter probe; (bottom row, left to right) ExG headset with additional sensor inputs; fNIR/EEG head set with coaxial EEG/fNIR sensor. B: AMI agent implementation for DACS utilizing multiple software agents to collect data from the RSS. It includes agents that manage experimental protocol administration and analysis algorithms for determination of ACS.

2.1 Software and Sensor Integration

All hardware sensors and software processing were synchronized through IHMC’s Adaptive Multiagent Integration (AMI) software framework [69]. Initially developed for the Defense Advanced Research Projects Agency’s AugCog program [1], AMI has matured into a high technology readiness level (TRL) over the past decade and is currently in preparation for open source release under the Apache 2.0 license open

source construct (Apache Software Foundation, Forrest Hills, MD). A large complement of devices, including video, audio, pressure, orientation and psychophysiological sensors, have already been integrated as software agents into this architecture. This inherently scalable cross-platform architecture, implemented in Java (Oracle Corp., Redwood City, CA) and C++, can use multiple displays/modalities simultaneously for real-time evaluation of various cognitive gauge display solutions. The AMI experiment controller agent allows a single operator to manage operation and data collection from multiple sensor devices, processing algorithms and displays, running on multiple computer nodes across wired, wireless or isolated networks with high precision time synchronization by applying sub-millisecond timestamps on each sensor data stream and through active measurement and correction of time base stability (milliseconds) between nodes.

2.2 Vocal Stress Analysis for Markers of Acute Cognitive Stress

The team analyzed a number of past aerospace mishaps with epochs of high stress that may have reached ACS levels, including United Airlines Flight 232 (a DC-10 that lost all hydraulics inflight but managed to reach Sioux City, IA, with more than two-thirds of the passengers and crew surviving [70]), an F15-C Eagle whose pilot died following a crash on takeoff (due to inadvertent cross-connected installation of the aircraft's flap control rods [71]), the Apollo XIII mishap [72], the US Airways Flight 1549 (call sign Cactus 1549, an Airbus A320 that made a controlled ditching in the Hudson River in New York City, following multiple bird-strikes and loss of both engines [73]), and an F16 training flight recovered by the automatic ground collision avoidance system (Auto-GCAS) after the student pilot experienced g-induced loss of consciousness (GLOC) potentially causing ACS in the instructor pilot [74]. Recordings from the latter three were analyzed to determine vocalization changes might serve as markers of ACS.

Methods. Audio recordings and transcripts from the latter three events (Apollo XIII, Cactus 1549 and F16 Auto-GCAS) have been distributed publicly and made available for analysis. Dramatizations of both Apollo XIII and Cactus 1549 were released, with actor Tom Hanks portraying the pilot in command (PIC) in both mishaps, albeit 21 years apart. This enabled evaluation of real and mimicked vocal stress from both the actual voice recordings and theatrical depictions. Audio files from the actual flights were downloaded from public sources in WAV or MPG format. Audio surrounding the mishaps extracted to WAV format files from purchased digital versatile disk (DVD) copies of the 1995 release of Apollo 13 (Universal Pictures, Inc., Universal City, CA) and the 2016 dramatization of Flight 1549 (Sully, Flashlight Films, LLC, Los Angeles, CA). The open-source, multitrack, audio editing software package Audacity[®] (version 2.0.6, <https://www.audacityteam.org>) was used to isolate each spoken utterance from the relevant events. The utterances were manually clipped to minimize extraneous noises (e.g., cockpit sounds, push-to-talk clicks, crew member utterances) and epochs without vocalizations. AMI agents were coded to determine speech cadence (words per second) and to calculate F0 for each of the events [45]. These AMI agents removed the ten leading and trailing data points from each utterance to remove transients prior to calculation of F0. While F0 magnitude lacks utility, due to

variations in individual baseline and background noises, increases can indicate increased stress. The AMI agent calculated cadence (words per second) using manually determined word counts for each utterance (verified against the official transcripts for Apollo XIII and Cactus 1459).

Results. All three of the events with publicly available audio were analyzed for both changes in speech cadence (words uttered per second) and for fundamental frequency (F0). For the F16 Auto-GCAS event, only the instructor pilot’s voice is available; however, for Apollo XIII and Cactus 1549, both the actual (flight) and the movie audio were used to identify if a trained actor could accurately mimic the changes in vocalization associated with a potential ACS event. In addition, Apollo XIII commander (CDR) Jim Lovell, command module pilot (CMP) Jack Swigert and lunar module pilot (LMP) all spoke during their event, as did all three of the actors who portrayed the crew (CDR-Tom Hanks, CMP-Kevin Bacon and LMP-Bill Paxton). The analysis excluded the SMP and LMP utterances, as well as the mission control center spacecraft capsule communicator (CAPCOM) during the event, astronaut Jack Lousma. While Co-pilot Jeffrey Skiles and air traffic controller Patrick Harten also appear in the Cactus 1549 flight audio, that analysis was likewise limited to the PIC. Because short utterances can provide unreliable results in this type of analysis, single word vocalizations were excluded.

F16 Auto-GCAS. The United States Air Force (USAF) F16 near mishap occurred during basic flight maneuvers on May 5, 2016 near Tucson, AZ, and fatalities were prevented because the automatic ground collision avoidance system (Auto-GCAS) recovered the aircraft. During the event, the instructor pilot called to incapacitated student multiple times (Table 1), repeating “Two, Recover” three times with increasing urgency, and a fourth time with less emphasis as the student’s F16 initiated the automated recovery. The instructor’s speech cadence (see Fig. 2) dropped for three successive utterances, as he extended the duration of each word and spoke more loudly to make his vocalization more salient. His cadence picked up again while the aircraft recovers. The F0 of his voice followed an opposite cycle, rising (a marker of cognitive stress) during the event and falling during the recovery. This provided a straightforward baseline analysis as the recording consisted of a single speaker sequentially uttering the same phrase followed by slightly longer phrases.

Table 1. F16 Instructor pilot utterances during and after Auto-GCAS event.

Utterance #	Word count	Condition	F16 Auto-GCAS, instructor pilot utterances
1	2	Event	Two, Recover
2	2	Event	Two, Recover!
3	2	Event	Two, Recover!!
4	2	Recovery	Two, Recover
5	7	Post-event	Two, get yourself back above the floor
6	6	Post-event	Alright, Two, climb back above...

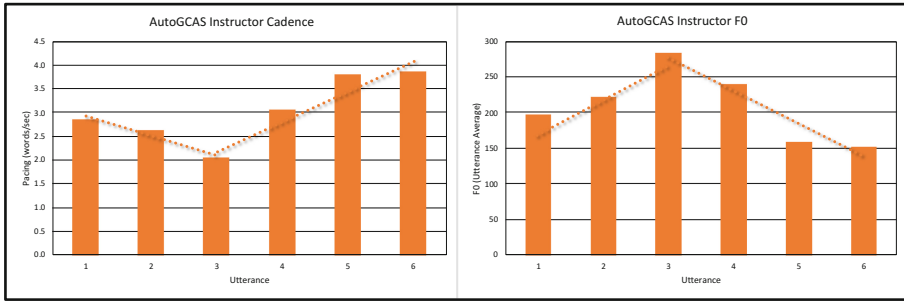


Fig. 2. F16 Auto-GCAS event instructor voice analysis with trend lines dividing utterances between event (1–3) and recovery (4–6) depicting, *left*, fall and rise in cadence and, *right*, rise and fall in F0 before and after the Auto-GCAS recovers the student’s aircraft.

Apollo XIII Oxygen Tank Mishap. Both the actual flight radio communications between the Apollo XIII crew and mission control recorded from the April 1970 mission [75] and from the movie, *Apollo 13*, were analyzed. The dialogue was first processed starting with the live television transmission event, which immediately preceded the oxygen (O₂) tank stir that triggered the explosion at mission elapsed time (MET) 55 h and 55 min, through to the observation of gas venting from the spacecraft approximately fifteen minutes later (Table 2). While the screenplay often differed (script, content, speaker, etc.) from the flight transcripts, much of the dialogue matched either verbatim or semantically (Table 2). The trendlines (see Fig. 3) suggest that both actual and movie CDR pacing, averaged within each utterance, spiked before the O₂ tank explosion, but just following an intentional cycling of a re-pressurization valve by LMP Haise; a prank that startled the CDR. The cadence for both the flight and the movie increases with utterance number 3, “Houston, we’ve had a problem” but then decreases throughout the course of the event, peaking again at utterance nine upon realizing that the spacecraft is venting a gas. For F0, also averaged within each utterance, the trendlines for both CDRs show a steep drop at utterance three, but Lovell shows a steady decrease until utterance nine, while Hanks shows a general increase in F0. This is evident to the casual listener because the actual CDR sounds more

Table 2. Apollo XIII (flight) and Apollo 13 (movie) O₂ tank explosion relevant utterances.

Utterance #	Word count	Condition	Flight (J. Lovell)	Word count	Movie (T. Hanks)
1	29	Pre-Event	Now, that round, uh, bag that’s just behind Fred that holds our, uh, vacuum hose and when get back inside the LM we’ll hook the vacuum off our suits	24	Now, when we get ready to land on the moon, Fred Haise and I will float through this access tunnel into the Lunar Module

(continued)

Table 2. (continued)

Utterance #	Word count	Condition	Flight (J. Lovell)	Word count	Movie (T. Hanks)
2	11	Pre-Event	Every time he does that our hearts- our hearts jump in our mouths	18	Fred Haise on the cabin repress valve. He really gets our hearts going every time with
3	6	Event	And, Houston, we've had a problem	5	Houston, we have a problem
4	4	Event	Main bus B undervolt	4	Main bus B undervolt
5	8	Event	We had a restart on our computer	5	We've got computer restart
6	6	Event	We had a PNGCS light, and uh	5	We've got a PNGCS light
7	4	Event	And the restart	5	Got a reset and a restart
8	19	Event	We are venting something...	17	We're venting something
9	6	Event	It's a gas of some sort	4	Gas of some sort

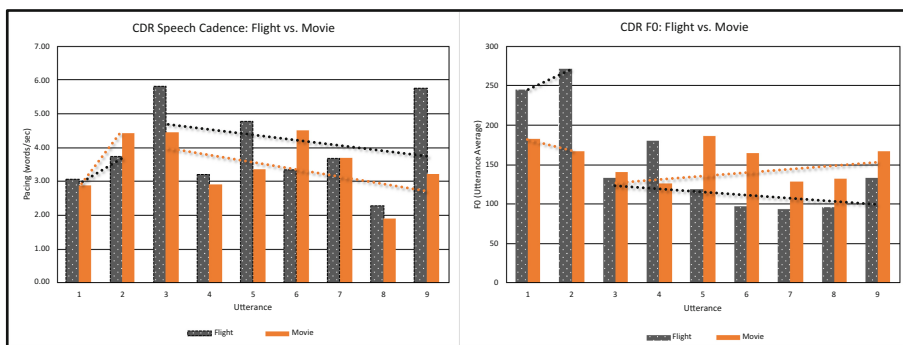


Fig. 3. Vocalization analysis from the Apollo XIII flight and the Apollo 13 movie. Speech cadence, *left*, and F0, *right*, by utterance, with trendlines for the CDRs (Lovell and Hanks).

matter-of-fact than his theatrical counterpart indicating, perhaps, a choice by Mr. Hanks to increase the drama of the event. An additional comparison between two Apollo XIII maneuvering events provides some additional insight into factors that may contribute to ACS (Table 3). As the movie depiction and the flight recordings did not have matching dialogue segments with useable audio, only the flight (CDR Lovell) utterances were used in this analysis.

Table 3. Pre and post mishap utterances associated with course/attitude correction burns.

Utterance #	Word count	Condition	Midcourse-2 correction Burn	Word count	Condition	Drift correction burn
1	7	Pre-Event	Okay, we'll do the gimbal test option	16	Post-Event	Well, we're, we're ATT hold for one thing—I mean, we're at minimum impulse
2	16	Pre-Event	Yes. We can hear and feel the, the engine gimbal as we do the test	15	Post-Event	I, I, Every time I try to, uh, I can't take that doggone roll out
3	7	Pre-Event	F-D-A-I scale 5, 5?	9	Post-Event	I got to wait until they get around to the bellyband
4	11	Pre-Event	A-S 58, we want Delta-V thrust A to normal	4	Post-Event	Okay. We'll try that
5	28	Pre-Event	Translation hand controller armed. Arm your rotational hand controller. I've already got mine armed. Okay, Fred...	15	Post-Event	Let me get around it, let's roll - let me, let it roll all the way
6	5	Pre-Event	Standing by for enter enable	7	Post-Event	I know, I know, but I mean-

Both segments analyzed relate to maneuvering events required to keep the spacecraft on course; the first occurred before the tank explosion (at MET 30 h and 38–40 min) and the second occurred after it (at MET 59 h and 05–06 min). The CDR's F0 (see Fig. 4) remains elevated throughout the process of learning a new skill, controlling the attitude of the combined stack of the Command and Service Module (CSM) and the Lunar Module (LM) from the LM rather than the CSM. The vocalization analysis indicates that managing the drift correction resulted in higher stress levels than responding to the O₂ tank explosion event itself.

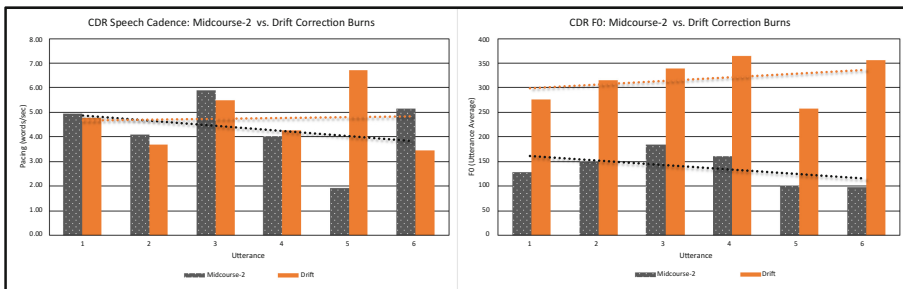


Fig. 4. Speech cadence, *left*, and F0, *right*, by utterance, with trendlines for the CDR (Lovell) during engine maneuvering burns before and after the O₂ tank explosion. The drift correction required learning to use controls with transposed responses due to the difference in CSM and LM control input strategies. This stress manifests with a higher and increasing trend in F0.

US Airways Flight 1549. Publically released audio communications between the PIC of US Airways Flight (Cactus) 1549 (Table 4) and the air traffic controller (TRACON) were analyzed along with the equivalent utterances from the depiction in the movie beginning with the bird strike event and continuing through the final communications prior to landing in the Hudson River (New York, NY). US Airways inherited the “Cactus” call sign used throughout the recordings when it merged with America West Airlines (Phoenix, AZ) in 2005. The PIC faced potential imminent death (for those on the aircraft and, likely, on the ground in a densely populated city) and a highly compressed decision-making time frame, while flying an aircraft with suddenly unpredictable responses, all of which likely caused him to increase his speech cadence and F0 as well as misspeak in his first call to TRACON (identifying as Cactus 1539, instead of 1549). Unlike the Apollo XIII/Apollo 13 analysis, in both the actual flight recording and movie audio, the PIC demonstrated an expected increase in pacing throughout the event (see Fig. 5). The charts below show a similarly close tracking between both recordings as the average F0 for each utterance increases throughout the evolution of the mishap. These features manifested despite the fact that the PIC maintained outwardly calm, deliberate and matter-of-fact vocalizations during communications with the TRACON as the mishap moved from declaration of emergency, through troubleshooting and discussion of alternatives to final transmission before the controlled ditching in the Hudson River. In contrast, in Apollo XIII O₂ event the crew

Table 4. Analysis of Cactus 1549 flight and movie utterances.

Utterance #	Word count	Condition	Flight (C. Sullenberger)	Word count	Movie (T. Hanks)
1	21	Event	Uh, this is, uh, Cactus 15-39. Hit birds, we've lost thrust (in/on) both engines. We are turning back towards LaGuardia	20	This is, uh, Cactus 15-49. Hit birds, we've lost thrust on both engines. We are turning back towards LaGuardia
2	3	Event	Two, Two, Zero	3	Two, Two, Zero
3	9	Event	We're unable. We may end up in the Hudson	10	We are unable. We may end up in the Hudson
4	19	Event	I'm not sure we can make any runway. Uh, what's over to our right anything in New Jersey maybe Teterboro?	21	I don't think we can make any runway. Uh, what about over to our right anything in New Jersey maybe Teterboro?
5	4	Event	We can't do it	4	We can't make it
6	6	Event	We're gonna be in the Hudson	7	We're gonna end up in the Hudson

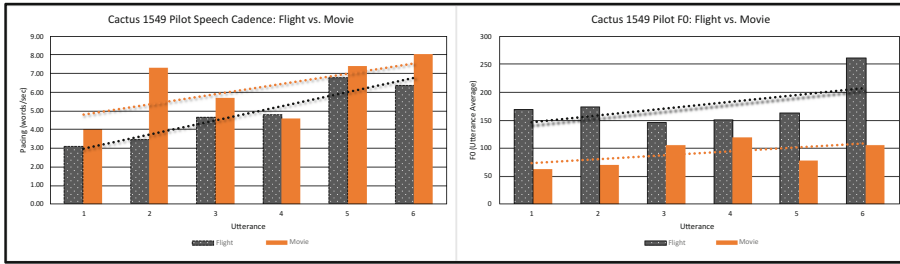


Fig. 5. Speech cadence, *left*, and F0, *right*, by utterance, with trendlines for the PIC for the actual flight (Sullenberger) and the theatrical version (Hanks) from the initial emergency call to TRACON through the last radio transmission. Both the flight and movie PICs showed similar trends in pacing and F0, indicating increasing PIC stress accurately mimicked in the movie.

worked the troubleshooting phase with less urgency until realizing that the spacecraft was venting a mission critical gas.

The TRACON’s verbal interactions with the PIC of Cactus 1549 were analyzed for comparison. Because the TRACON made many more utterances than the PIC, pacing and F0 were categorized and averaged by phase of the mishap. The TRACON showed increasing cadence (indicating a shift to an increased cognitive stress level) as he became aware of the situation (“Informed”) that stabilized as he pursued options, lost contact with the pilot and eventually faced “(Perceived) Failure” when he was removed from his duties by his supervisor. Even though he demonstrated only a modest increase after the bird strike, this analysis serves as positive evidence that cognitive stress can be affected without rising to the level of imminent danger to the speaker’s life (Fig. 6).

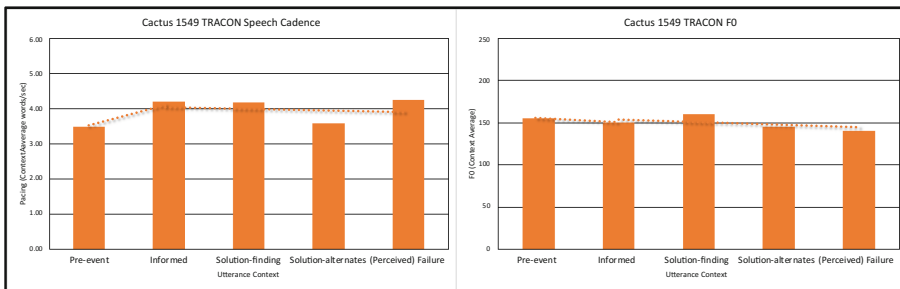


Fig. 6. Speech cadence, *left*, and F0, *right*, by utterance, with trendlines for the TRACON response through the mishap phases. Pacing increases initially while F0 remains stable throughout.

3 Discussion

These examples indicate that signals of cognitive stress and strain can be extracted from analysis of vocalizations. The combination of speech cadence and F0 metrics would likely provide better reliability than either one alone. The analysis, however,

also indicates that individuals can mimic some elements of vocal ACS markers. Therefore, vocalization ACS markers should provide additional sensor and contextual information to a more multisensory approach as proposed for the RSS. In all these events, an unexpected change in response to input controls caused rapid increases in stress and likely ACS. The PIC of Cactus 1549, Capt. Chesley Sullenberger described the event as: "... the most dire emergency of my career. It was one of extreme workload, beyond extreme time compression. We had only seconds to decide and minutes to find a course of action and execute it." [73] Hallmarks of ACS provocative events likely would include rapid increase in baseline workload, unpredictable response to prior learned behavior, time pressure and high cost of failure (due to error or inaction). The publically available audio files utilized in the above analysis were recorded with varying levels of microphone quality, digitization and compression. In addition, the actual flight recordings contained other cockpit noises, vocalizations by other crewmembers and ground controllers, radio transmission artifacts and other contamination. The movie audio also included musical scoring and modulation effects used to mimic the midrange bandpass filtering associated with radio communications. However, this preliminary analysis sought to identify changes in vocalizations within each individual's recordings that could indicate ACS. This analysis indicates that pacing can increase or decrease in response to increased stress, depending on the content of the utterance while F0 tends to increase. This suggests that the DACS concept may need to identify changes in deviation from baseline cadence or understand the content of the utterances.

4 Future Work

The algorithms used for processing the audio have been ported from an offline process into AMI agents that can process streaming audio data in real-time. Combining natural language processing (NLP) with the cadence and F0 analysis would enable extraction of semantic content that would provide additional contextual cues for the DACS. To improve the quality of the audio, we propose to use a throat microphone that would minimize the effect of ambient noise. We have designed and fabricated a low noise, multi-stage audio amplifier printed circuit board (PCB) to acquire and digitize audio from the throat microphone. This circuit, optimized for flat frequency response in the range of human F0, could be further reduced in size to mount directly with the microphone. Incorporating the vocalization analysis with the other discussed psychophysiologic sensors (to be described in a future publication) would allow the RSS to provide a comprehensive, multisensory laboratory tool for developing and optimizing the subset of sensor technologies needed for a wearable DACS system for use in operational settings, including aviation environments.

In order to achieve this goal, the research team developed an operationally relevant aviation testbed to evaluate sensor response to and machine learning approaches for detecting ACS. This evaluation platform consists of a fixed base flight simulator, with both programmable workload task levels and system failure modes. The aviation mishaps that were analyzed showed that unexpected change in the vehicle performance, control response and stability, tended to increase stress more so than the

cognitive process of identifying the underlying problem (e.g., the Apollo attitude correction burns increased F0 more than the mishap itself). The proposed simulation delivers a moderate, but manageable baseline workload (e.g., flying an hour long cross-country mission with multiple waypoints and heading changes) while managing aircraft subsystems operating nominally. This implementation uses an F35 Lightning II motion model (AOA Simulations, Redmond, WA) running in X-Plane (v11, Lamina Research, Columbia, SC) using a joystick, throttle (side stick) and rudder pedals (Thrustmaster Warthog HOTAS, Guillemot Corp., La Gacilly, France). Adding an intermittent control input failure (i.e., reversal or variable lag) during high workload or high stress should provoke ACS in a motivated and engaged participant (Fig. 7).



Fig. 7. A: DACS cockpit simulator example during ACS provoking event. During a standard-rate left turn, the pitch response to the control stick input reversed intermittently. This caused unpredictable aircraft behavior and pilot induced oscillation (PIO). **B:** *left*, Peltor throat microphone and, *right*, initial prototype low noise multistage amplifier and digitizer PCB.

This ACS testing platform embeds a standard alphanumeric N-back memory task [55] into the flight instrument displays, which allows the simulation to increase cognitive workload (using this secondary task) without altering the primary task. Increasing N from one (“Is the currently displayed character the same as the one displayed just before it?”) to, for example, three (“Is the currently displayed character the same as the one displayed three characters before?”) can dramatically increase the cognitive workload. Decreasing the time allowed for a response and the duration of display of the target character, as well as negative auditory feedback (i.e., a noxious buzzer following each incorrect response) can increase time pressure and cognitive stress. To improve the system performance, we have recently implemented an AMI agent that automatically determines the start and end of and the number of words in each utterance, to enable automated vocalization analysis.

We propose to evaluate the RSS and develop the DACS with student aviators, a population with high interest in the aviation task, but without significant experience managing cockpit emergencies. Increasing cockpit workload and stress with the embedded N-back test and the intermittent/variable control response failures, should provoke ACS by making a previously manageable task both unpredictable and impossible. Initial integration evaluations with research staff revealed that even when

the failure type and onset was known, control rapidly degenerated into pilot induced oscillation (PIO). Returning the control to normal functionality for the majority of the mission should prevent the participants from losing interest (engagement) in the task altogether.

5 Conclusions

The proposed psychophysiologic sensing modalities each provide insight into individual states of cognitive stress and strain levels. Their integration into the multisensory RSS via the multiagent AMI framework provides a real-time mechanism for test and evaluation of each sensor modality, both individually and collectively through the use of statistical, machine learning and control theoretic algorithmic analyses. Successful implementation of the RSS in the proposed testbed would determine which RSS sensors provide the most robust and reliable markers inclusion in a miniaturized and wearable DACS system for aviation. Many researchers have already demonstrated that a single sensor can detect multiple modalities (e.g., detection of eye movements from EEG), likewise, multiple sensors can also be integrated to occupy the same physical space (e.g., QUASAR's fNIR/EEG nodes, incorporation of LEMG into the throat microphone structure). Therefore, the resulting DACS system could become unobtrusive with the low size, weight and power requirements (SWaP) required for integration into aviation and other operational domains. This would drive the development and use of ACS detection as an input for training feedback, performance evaluation, and engagement of adaptive automation. Thus, an operational DACS system could potentially improve skill acquisition by mitigating student frustration when learning a new task, evaluate skill application proficiency during evaluation (in simulated or actual operations) and provide a key input for task allocation adjustment in mixed human-automation teams.

Acknowledgement. This work was supported by the Defense Advanced Research Projects Agency Phase I Small Business Technology Transfer program topic number D16C-003: Optimizing Human-Automation Team Workload through a Non-Invasive Detection System under award number DA17PC00172 Detector of Acute Cognitive Strain (DACS).

References

1. Schmorrow, D.D., Kruse, A.A.: Augmented cognition. In: Bainbridge, W.S. (ed.) *Berkshire Encyclopedia of Human-Computer Interaction*, pp. 54–59. Berkshire Publishing Group, Great Barrington (2004)
2. McDonald, N.J., Soussou, W.: QUASAR's QStates cognitive gauge performance in the cognitive state assessment competition 2011. In: 33rd Annual International Conference of the IEEE EMBS. Boston, Massachusetts USA, 30 August–3 September, pp. 6542–6546 (2011)
3. Ceballos, N.A., Giuliano, R.J., Wicha, N.Y., Graham, R.: Acute stress and event-related potential correlates of attention to alcohol images in social drinkers. *J. Stud. Alcohol Drugs* **73**(5), 761–771 (2012)

4. Covey, T.J., Shucard, J.L., Violanti, J.M., Lee, J., Shucard, D.W.: The effects of exposure to traumatic stressors on inhibitory control in police officers: a dense electrode array study using a Go/NoGo continuous performance task. *Int. J. Psychophysiol.* **87**(3), 363–375 (2013)
5. Dierolf, A.M., Fechtner, J., Bohnke, R., Wolf, O.T., Naumann, E.: Influence of acute stress on response inhibition in healthy men: an ERP study. *Psychophysiology* **54**(5), 684–695 (2017)
6. Banis, S., Geerligs, L., Lorist, M.M.: Acute stress modulates feedback processing in men and women: differential effects on the feedback-related negativity and theta and beta power. *PLoS ONE* **9**(4), e95690 (2014)
7. Compton, R.J., Hofheimer, J., Kazinka, R.: Stress regulation and cognitive control: evidence relating cortisol reactivity and neural responses to errors. *Cogn. Affect. Behav. Neurosci.* **13**(1), 152–163 (2013)
8. Singh, Y., Sharma, R.: Individual alpha frequency (IAF) based quantitative EEG correlates of psychological stress. *Indian J. Physiol. Pharmacol.* **59**(4), 414–421 (2015)
9. Goodman, R.N., Rietschel, J.C., Lo, L.C., Costanzo, M.E., Hatfield, B.D.: Stress, emotion regulation and cognitive performance: the predictive contributions of trait and state relative frontal EEG alpha asymmetry. *Int. J. Psychophysiol.* **87**(2), 115–123 (2013)
10. Subhani, A.R., Xia, L., Malik, A. S., Othman, Z.: Quantification of physiological disparities and task performance in stress and control conditions. In: 2013 35th Annual Conference of the IEEE Engineering in Medicine and Biology Society, Osaka, Japan, pp. 2060–2063 (2013)
11. Subhani, A.R., Likun, X., Saeed Malik, A.: Association of autonomic nervous system and EEG scalp potential during playing 2D Grand Turismo 5. In: 2012 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, pp. 3420–3423 (2012)
12. Riera, A., Soria-Frisch, A., Albajes-Eizagirre, A., Cipresso, P., Grau, C., Dunne, S., Ruffini, G.: Electro-physiological data fusion for stress detection. *Stud. Health Technol. Inf.* **181**, 228–232 (2012)
13. Putman, P., Arias-Garcia, E., Pantazi, I., van Schie, C.: Emotional Stroop interference for threatening words is related to reduced EEG delta-beta coupling and low attentional control. *Int. J. Psychophysiol.* **84**(2), 194–200 (2012)
14. Kawasaki, S., Nishimura, Y., Takizawa, R., Koike, S., Kinoshita, A., Satomura, Y., Kasai, K., et al.: Using social epidemiology and neuroscience to explore the relationship between job stress and frontotemporal cortex activity among workers. *Soc. Neurosci.* **10**(3), 230–242 (2015)
15. Ogata, H., Mukai, T., Yagi, T.: A study on the frontal cortex in cognitive tasks using near-infrared spectroscopy. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, pp. 4731–4734 (2007)
16. Ishii, Y., Ogata, H., Takano, H., Ohnishi, H., Mukai, T., Yagi, T.: Study on mental stress using near-infrared spectroscopy, electroencephalography, and peripheral arterial tonometry. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, Canada, pp. 4992–4995 (2008)
17. Shi, J., Sakatani, K., Okamoto, M., Yamaguchi, Y., Zuo, H.C.: Correlation between LIFG and autonomic activation during stressful tasks: a functional near-infrared spectroscopy (fNIRS) study. *J. Huazhong Univ. Sci. Technol. [Med. Sci.]* **34**(5), 663–671 (2014)
18. Takizawa, R., Nishimura, Y., Yamasue, H., Kasai, K.: Anxiety and performance: the disparate roles of prefrontal subregions under maintained psychological stress. *Cereb. Cortex* **24**(7), 1858–1866 (2014)

19. Tanida, M., Katsuyama, M., Sakatani, K.: Relation between mental stress-induced prefrontal cortex activity and skin conditions: a near-infrared spectroscopy study. *Brain Res.* **1184**, 210–216 (2007)
20. Tanida, M., Katsuyama, M., Sakatani, K.: Effects of fragrance administration on stress-induced prefrontal cortex activity and sebum secretion in the facial skin. *Neurosci. Lett.* **432**(2), 157–161 (2008)
21. Krantz, G., Forsman, M., Lundberg, U.: Consistency in physiological stress responses and electromyographic activity during induced stress exposure in women and men. *Integr. Physiol. Behav. Sci.* **39**(2), 105–118 (2004)
22. Nilsen, K.B., Sand, T., Stovner, L.J., Leistad, R.B., Westgaard, R.H.: Autonomic and muscular responses and recovery to one-hour laboratory mental stress in healthy subjects. *BMC Musculoskelet. Disord.* **8**, 81 (2007)
23. Schleifer, L.M., Spalding, T.W., Kerick, S.E., Cram, J.R., Ley, R., Hatfield, B.D.: Mental stress and trapezius muscle activation under psychomotor challenge: a focus on EMG gaps during computer work. *Psychophysiology* **45**(3), 356–365 (2008)
24. Shahidi, B., Haight, A., Maluf, K.: Differential effects of mental concentration and acute psychosocial stress on cervical muscle activity and posture. *J. Electromyogr. Kinesiol.* **23**(5), 1082–1089 (2013)
25. Taib, M.F., Bahn, S., Yun, M.H.: the effect of psychosocial stress on muscle activity during computer work: comparative study between desktop computer and mobile computing products. *Work* **54**(3), 543–555 (2016)
26. Marker, R.J., Campeau, S., Maluf, K.S.: Psychosocial stress alters the strength of reticulospinal input to the human upper trapezius. *J. Neurophysiol.* **117**(1), 457–466 (2017)
27. Kristiansen, J., Mathiesen, L., Nielsen, P.K., Hansen, A.M., Shibuya, H., Petersen, H.M., Sogaard, K., et al.: Stress reactions to cognitively demanding tasks and open-plan office noise. *Int. Arch. Occup. Environ. Health* **82**(5), 631–641 (2009)
28. Larsman, P., Thorn, S., Sogaard, K., Sandsjo, L., Sjogaard, G., Kadefors, R.: Work related perceived stress and muscle activity during standardized computer work among female computer users. *Work* **32**(2), 189–199 (2009)
29. Luijckx, R., Hermens, H.J., Bodar, L., Vossen, C.J., Van Os, J., Lousberg, R.: Experimentally induced stress validated by EMG activity. *PLoS ONE* **9**(4), e95215 (2014)
30. Dietrich, M., Verdolini Abbott, K.: Psychobiological stress reactivity and personality in persons with high and low stressor-induced extralaryngeal reactivity. *J. Speech Lang. Hear. Res.* **57**(6), 2076–2089 (2014)
31. Helou, L.B., Wang, W., Ashmore, R.C., Rosen, C.A., Verdolini Abbott, K.: Intrinsic laryngeal muscle activity in response to autonomic nervous system activation. *Laryngoscope* **123**(11), 2756–2765 (2013)
32. Dietrich, M., Verdolini Abbott, K.: Vocal function in introverts and extraverts during a psychological stress reactivity protocol. *J. Speech Lang. Hear. Res.* **55**(3), 973–987 (2012)
33. Sigmund, M., Prokes, A., Brabec, Z.: Statistical analysis of glottal pulses in speech under psychological stress. In: 16th European Signal Processing Conference, Lausanne, Switzerland, pp. 1–5. IEEE (2007)
34. Scherer, S., Hofmann, H., Lampmann, M., Pfeil, M., Rhinow, S., Schwenker, F., Palm, G.: Emotion recognition from speech: stress experiment. In: 6th International Conference on Language Resources and Evaluation (LREC), pp. 1325–1330 (2008)
35. Frampton, M., Sripada, S., Augusto, R., Bion, H., Peters, S.: Detection of time-pressure induced stress in speech via acoustic indicators. In: 11th Annual Meeting of the Special Interest Group of Discourse and Dialogue, Tokyo, Japan, pp. 253–256 (2010)

36. Casale, S., Russo, A., Serrano, S.: Multi-style classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Commun.* **49**(10), 801–810 (2007)
37. Hansen, J.H.L., Womack, B.: Feature analysis and neural network-based classification of speech under stress. *IEEE Trans. Speech Audio Process.* **4**(4), 307–313 (1996)
38. Patil, S.A., Hansen, J.H.L.: The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification. *Speech Commun.* **52**(4), 327–340 (2010)
39. Zhou, G.: Nonlinear feature based classification of speech under stress. *IEEE Trans. Speech Audio Process.* **9**(3), 201–216 (2001)
40. Protopapas, A., Liberman, P.: Fundamental frequency of phonation and perceived emotional stress. *J. Acoust. Soc. Am.* **101**(4), 2267–2277 (2001)
41. Brenner, M., Doherty, E.T., Shipp, T.: Speech measures indicating workload demand. *Aviat. Space Environ. Med.* **65**(1), 21–26 (1994)
42. Titze, I.R.: *Principles of Voice Production*. Prentice Hall, Englewood Cliffs (1994)
43. National Transportation Safety Board (NTSB): In-Flight Separation of Vertical Stabilizer, American Airlines Flight 587 Airbus Industrie A300-605R, N14053, Belle Harbor, New York, 12 November 2001, Washington, DC (2004)
44. Singh, S., Bucks, R., Cuerden, J.: Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech. *Aphasiology* **15**(6), 571–584 (2001)
45. Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., Kaye, J.A.: Spoken language derived measures for detecting mild cognitive impairment. *IEEE Trans. Audio Speech Lang. Process.* **19**(7), 2081–2090 (2011)
46. Junqua, J.: The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex. *Speech Commun.* **20**(1–2), 13–22 (1996)
47. Ikeno, I., Varadarajan, V., Patil, S., Hansen, J.H.L.: UT-Scope: speech under lombard effect and cognitive stress. In: 2007 IEEE Aerospace Conference, pp. 1–7. Big Sky, MT (2007)
48. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F.: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* **44**, 58–75 (2014)
49. Stepanek, J., Pradhan, G.N., Cocco, D., Smith, B.E., Bartlett, J., Studer, M., Cevette, M.J., et al.: Acute hypoxic hypoxia and isocapnic hypoxia effects on oculometric features. *Aviat. Space Environ. Med.* **85**(7), 700–707 (2014)
50. Veltman, J.A., Gaillard, A.W.: Physiological workload reactions to increasing levels of task difficulty. *Ergonomics* **41**(5), 656–669 (1998)
51. Tsai, Y.F., Viirre, E., Strychacz, C., Chase, B., Jung, T.P.: Task performance and eye activity: predicting behavior relating to cognitive workload. *Aviat. Space Environ. Med.* **78** (5 Suppl), B176–B185 (2007)
52. Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O.N., Atkins, M.S.: Workload assessment of surgeons: correlation between NASA TLX and blinks. *Surg. Endosc.* **26**(10), 2746–2750 (2012)
53. Macatee, R.J., Albanese, B.J., Schmidt, N.B., Cogle, J.R.: Attention bias towards negative emotional information and its relationship with daily worry in the context of acute stress: an eye-tracking study. *Behav. Res. Ther.* **90**, 96–110 (2017)
54. Dehais, F., Causse, M., Vachon, F., Tremblay, S.: Cognitive conflict in human-automation interactions: a psychophysiological study. *Appl. Ergon.* **43**(3), 588–595 (2012)
55. Kirchner, W.K.: Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* **55**(4), 352–358 (1958)

56. Mandrick, K., Peysakhovich, V., Remy, F., Lepron, E., Causse, M.: Neural and psychophysiological correlates of human performance under stress and high mental workload. *Biol. Psychol.* **121**(Pt A), 62–73 (2016)
57. Vinski, M.T., Watter, S.: Being a grump only makes things worse: a transactional account of acute stress on mind wandering. *Front. Psychol.* **4**, 730 (2013)
58. Henckens, M.J., Hermans, E.J., Pu, Z., Joels, M., Fernandez, G.: Stressed memories: how acute stress affects memory formation in humans. *J. Neurosci.* **29**(32), 10111–10119 (2009)
59. Ren, P., Barreto, A., Gao, Y., Adjouadi, M.: Comparison of the use of pupil diameter and galvanic skin response signals for affective assessment of computer users. *Biomed. Sci. Instrum.* **48**, 345–350 (2012)
60. Wang, X., Liu, B., Xie, L., Yu, X., Li, M., Zhang, J.: Cerebral and neural regulation of cardiovascular activity during mental stress. *BioMed. Eng. OnLine* **15**(Suppl 2), 160 (2016)
61. Myrtek, M., Weber, D., Brugner, G., Muller, W.: Occupational stress and strain of female students: results of physiological, behavioral, and psychological monitoring. *Biol. Psychol.* **42**(3), 379–391 (1996)
62. Visnovcova, Z., Mestanik, M., Javorka, M., Mokra, D., Gala, M., Jurko, A., Tonhajzerova, I., et al.: Complexity and time asymmetry of heart rate variability are altered in acute mental stress. *Physiol. Meas.* **35**(7), 1319–1334 (2014)
63. Tanev, G., Saadi, D.B., Hoppe, K., Sorensen, H.B.: Classification of acute stress using linear and non-linear heart rate variability analysis derived from sternal ECG. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, pp. 3386–3389 (2014)
64. Minakuchi, E., Ohnishi, E., Ohnishi, J., Sakamoto, S., Hori, M., Motomura, M., Kawaguchi, T., et al.: Evaluation of mental stress by physiological indices derived from finger plethysmography. *J. Physiol. Anthropol.* **32**(1), 17 (2013)
65. Ahlund, C., Pettersson, K., Lind, L.: Influence of different types of stressors on the waveform of the peripheral arterial pulse in humans. *Blood Press.* **12**(5–6), 291–297 (2003)
66. Guez, J., Saar-Ashkenazy, R., Keha, E., Tiferet-Dweck, C.: The effect of trier social stress test (TSST) on item and associative recognition of words and pictures in healthy participants. *Front. Psychol.* **7**, 507 (2016)
67. Winslow, B.D., Chadderdon, G.L., Dechmerowski, S.J., Jones, D.L., Kalkstein, S., Greene, J.L., Gehrman, P.: Development and clinical evaluation of an mHealth application for stress management. *Front. Psychiatry* **7**, 130 (2016)
68. Li, D., Winfield, D., Parkhurst, D.J.: Starburst: a hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In: 2005 IEEE Computer Society Conference Computer Vision and Pattern Recognition - Workshops, San Diego, CA, pp. 79–79 (2005)
69. Johnson, M.J., Kulkarni, S.P., Raj, A.K., Carff, R.W., Bradshaw, J.M.: AMI: an adaptive multi-agent framework for augmented cognition. In: 1st International Conference on Augmented Cognition. Lawrence Erlbaum Associates, Mahwah (2005)
70. National Transportation Safety Board (NTSB): Aircraft Accident Report—United Airlines Flight 232, McDonnell Douglas DC-10-10, Sioux Gateway Airport, Sioux City, Iowa, 19 July 1989, Report NO. NTSB/AAR-90/06. NTSB Bureau of Accident Investigation, Washington, DC (1990)
71. Thompson, M.: Placing blame at any cost. *TIME* **148**(25), 43–44 (1996)
72. Cortright, E.M.: Report of Apollo 13 review board. United States National Aeronautics and Space Administration (1970)

73. Hersman, D.A., Hart, C.A., Sumwalt, R.L.: Loss of Thrust in Both Engines After Encountering a Flock of Birds and Subsequent Ditching on the Hudson River, Accident Report NTSB/AAR-10/03. National Transportation Safety Board, Washington DC (2010)
74. Norris, G.: Auto-GCAS saves unconscious F-16 Pilot—declassified USAF footage. *Aviat. Week Space Technol.* (2016)
75. Apollo XIII Audio Collection. <https://archive.org/details/Apollo13Audio/>. Accessed 25 Jan 2018



Classification Procedure for Motor Imagery EEG Data

Ellton Sales Barros^(✉) and Nelson Neto

Faculty of Computer Science, Federal University of Pará, Belém, PA, Brazil
elltonsalesbarros@gmail.com, dnelsonneto@gmail.com

Abstract. Brain computer interface establishes a new model of communication, whereby it is possible to communicate using only cerebral signals, that can be obtained from different kind of cerebral stimuli. By the way, one of the most common stimulus is the motor imagery of the arms. However, since a set of variables leads to different levels of classification accuracy, it is necessary to search for procedures that can enhance the recognition accuracy of brain signals in order to create more precise systems. This paper proposes a classification procedure for discrimination of two motor imagery classes obtained using the Emotiv EPOC+ EEG signal acquisition device. The Emotiv EPOC+ has 14 input channels, but only four were used – the ones directly related with the capture of motor imagery signals. The presented procedure was created considering the MI common spatial pattern package from the OpenVibe software and the support vector machine (SVM) classification approach. As well, the procedure runs under the OpenVibe scenarios. A database with motor imagery signals from five subjects was built in order to perform the classification tests. In order to select the best features, several aspects from the signal acquisition until the classification process were analysed, such as selection of the best Kernel to SVM classifier, frequency band, filter output channels, and a grid-search to estimate the classifier parameters. At the end, an increase of 28,96% in the mean accuracy was achieved, regarding to the OpenVibe MI standard scenario.

Keywords: Brain computer interface · Support vector machine
Motor imagery

1 Introduction

Brain computer interface (BCI) is an alternative communication system that allows a user to interact with the environment through the conversion of brain signals into commands without the use of the neuromuscular system [1, 2]. A BCI system aims to offer people with disabilities the possibility of communication, controlling computers and other devices in an independent way, improving their quality of life, and reducing social cost [2, 3].

A crucial point within a BCI system is the way in which it extracts the electroencephalogram (EEG) signal features for better accuracy in signal recognition, since recognition accuracy affects system performance [4]. It is known that a set of variables leads to different levels of classification accuracy, such as positions of electrodes and

number of channels [2]. Another well-known challenge is to extract discriminative characteristics of raw EEG signals, since this method presents a poor signal-to-noise ratio and a mixture of different sources of brain activities [5]. Other characteristics that have great relevance in the accuracy of signal identification are the selection of the frequency band and the time segment of the EEG data [5].

In a BCI system, motor imagery (MI) refers to the signals obtained while the user is imagining a motor action [3]. When a subject imagines a limb movement, specific frequencies change, allowing EEG signals to be used to control a BCI system [6]. The identification of MI is performed by classifying the activity of power attenuation of sensorimotor rhythms, known as event-related desynchronization (ERD), and an increase of power of sensorimotor rhythms, known as event-related synchronization (ERS) [7]. ERD/S events normally occur in the Mu and Beta frequency bands around 8–32 Hz [8]. In this way, these signals can be converted into computer commands.

As explained, the difficulties of creating a MI-based BCI system are directly related to the choice of characteristics that optimize the signal classification performed by a computer. So, in this context, this work aims at developing an MI classification procedure using a EEG signal acquisition device, called Emotiv EPOC+. The proposed procedure is based on the well-known MI common spatial pattern (CSP) software package, part of the OpenVibe signal processing and classification toolkit [9]. During the process of creating the procedure, initial experiments were made in order to choose the kernel of the classifier that would be used. Then, considering the chosen classifier, the analyses were extended to select the best frequency band, filter parameters, and classifier parameters. Considering the mean accuracy of related works, we assumed a value of recognition accuracy equal or greater than 80% to be satisfactory.

The remainder of the paper is organized as follows. Section 2 presents the related works. Section 3 describes the materials and methods used in this work. A summary of the results is presented and discussed in Sect. 4. Finally, Sect. 5 summarizes our conclusions and addresses future works.

2 Related Works

This section addresses only works that performed the classification of imagined movements and also used the Emotiv EPOC+ device.

In [3], a classification of four MI (forward, backward, left, and right) was performed. The authors used wavelet daubechies and symlets for feature extraction, and multilayer perceptron, simple logistic, and bagging for classification. The signals were acquired from five healthy men aged between 26–35 years. The subjects imagined 40 s for each of the classes to be classified. The best accuracy was obtained with wavelet daubechies and simple logistic, with a mean total accuracy of 80.4% using a 10 fold cross-validation.

The identification of three mental states of MI was performed in the work presented in [10]. The considered states were: rest, up, and down, with a capture time of 60, 10, and 10 s, respectively. The EEG signals were captured from only eight channels, the characteristic extraction was based on the power spectrum, and the classifier used was an artificial neural network with a backpropagation learning algorithm. The algorithm

presented an accuracy of 72% in online tests performed with three subjects, with a training time of 15 min for each of them.

In [11], the authors present a time-series discrimination methodology called motor imagery discrimination by relevance analysis (MIDRA) to support BCI systems. The classifier used was a K-Nearest Neighbor, with 10 folds in cross-validation. The signal captured during the MI lasted four seconds, and each subject performed 160 trials: 80 MI related to the right hand and 80 MI related to the left hand. The classification accuracy values for the two subjects considered were 92.5% and 90%.

It is noteworthy that none of the related studies performed a complete analysis of the variables that make up the entire EEG signal classification process. In this context, this work intends to investigate how far a complete evaluation of the characteristics from acquisition until classification could improve the accuracy of a BCI system based on imagined movements.

3 Materials and Methods

3.1 Experimental Protocol

A dataset was created in order to investigate the classification accuracy of different MI tasks. This dataset has MI data from five male subjects, right-handed, aged between 20–25 years, and with no prior BCI experience. In total, each subject has performed 24 MI trials, 12 to the right arm and 12 to the left arm, given a total of 120 trials. During the MI period, the subject was asked to imagine the constant movement of the arm, according to the stimulus presented previously, right or left, until the end of the experiment, which was indicated on the screen.

Regarding the capture session protocol, each volunteer was placed comfortably in a chair in front of a computer display where the simulation experiment occurred. The subject was asked to be relaxed with his hands on his knees and physical movement was not allowed during the MI sessions. The performance of the subject and the screen were completely recorded during the experiment, for later evaluation in case of inconsistency or incoherence of the data. The time between trials was randomly chosen (between 1.5–3.5 s), as well as the stimuli order (right or left), therefore, the subject could not predict when the stimulation starts, avoiding the creation of a synchronized pattern, making it difficult the classification. The goal was performing self-paced sessions. Figure 1 shows the characteristics of a trial.



Fig. 1. Scheme of the MI capture session protocol. Notice that only three seconds of each trial were used in the classification process.

The experimental scenario was built by using the OpenVibe software [5]. It is a platform that allows the acquisition, filtering, processing, and classification of brain signals and fast prototyping of BCI systems. For the sake of experimentation, we explored three OpenVibe scenarios in this work: (1) Acquisition scenario: the EEG data was acquired and stored in a file; (2) CSP training scenario: the common spatial filter (CSP) was trained; and (3) Training and classification scenario: the classifier was trained and tested.

3.2 Data Acquisition

The EEG data was acquired by using an Emotiv EPOC+ neuroheadset with 14 electrodes placed according to the international 10–20 locations [12] and sampled at 128 Hz. EEG signals were filtered using a zero-phase bandpass Butterworth 4th order filter with cutoff frequencies in the range of 5–30 Hz. After the acquisition, as MI is related to motor actions, we used only channels located in the frontal lobe {F3, F4} and fronto-central lobe {FC5, FC6} that capture activities from premotor and motor cortex [13].

3.3 Methods

Figure 2 shows the adopted classification methodology. The band power is a function that returns the power average for a given input signal. First, the signal is squared becoming smoothed [14], and after the band power, the Eq. 1 is applied to the power average, as it assists in the improvement of the classification performance.

$$\ln(x + 1) \quad (1)$$

For classification, we used the support vector machine (SVM) approach, since it is fast and has shown good results in EEG analysis [4, 15–17]. Moreover, SVM has been shown to be a good binary classification algorithm with an excellent generalization capability [18]. After that, a study on the accuracy of the SVM with different kernels was performed in order to choose the one with the best result. The next evaluation was performed to investigate which frequency bands improve classification accuracy. Finally, the number of output channels of the CSP filter was determined, the SVM parameters were analysed and defined, and the classification results were obtained.



Fig. 2. Adopted classification methodology (signal = x).

4 Results and Discussions

In order to select the type (Nu-SVM or C-SVM) and the kernel of the SVM, the 24 MI trials dataset was used to run on different kernels, as shown in Fig. 3, and the mean accuracy was calculated. As can be seen, Nu-SVM with radial basis function

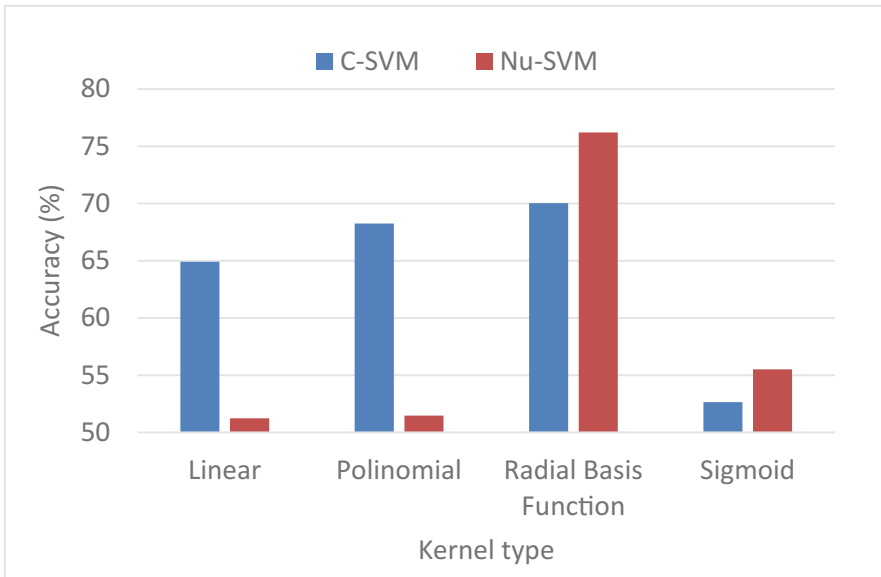


Fig. 3. Mean accuracy (%) for C-SVM, Nu-SVM, and different kernels.

(RBF) kernel was selected because of the best classification result. The cross-validation considered 5 folds, and this number of folds was maintained as standard for the next analyses.

4.1 Evaluating the Importance of Frequency Bands in Classification Accuracy

According to [19], the effectiveness of a BCI system depends on the choice of the frequency band. Then, in order to select the appropriate frequency band, an analysis was performed considering the frequency spectrum Mu (8–12 Hz) and Beta (12–30 Hz) which refer to frequencies related to MI [11]. The frequency band values were varied using a bandpass Butterworth filter and the mean accuracy was obtained.

As shown in Fig. 4, the final value of the frequency range was decreased, with the 8–10 Hz frequency band achieving the best result. Then, the lower frequency limit was changed, as shown in Fig. 5. The value of the frequency band of 9–10 Hz was the one that obtained the best classification accuracy. This is consistent with MI studies previously reported [2], where the highest accuracy was obtained for the 8–15 Hz frequency band. In this work, we decided to keep the range 9–10 Hz as the standard for the next evaluations.

4.2 Evaluating the Best Number of CSP Filter Output Channels

The CSP filter extracts characteristics of the signals that maximize the difference between mental states [14]. The filter can generate two or more output channels by

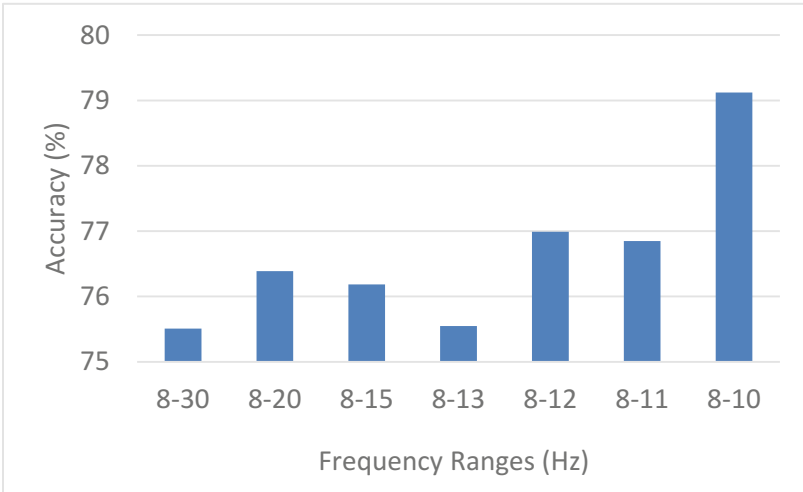


Fig. 4. Mean accuracy (%) decreasing the final value of the frequency range (8–30/8–10 Hz).

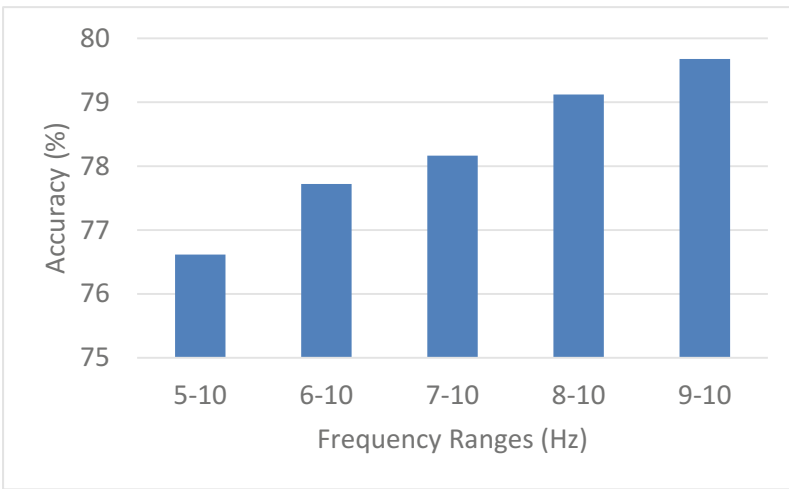


Fig. 5. Mean accuracy (%) increasing the initial value of the frequency range (5–10/9–10 Hz).

combining the input signals. The analysis was performed considering the mean accuracy and using CSP filters with 2 until 14 output channels, as can be seen in Fig. 6. The best value was obtained with six channels, so this number was maintained for the next experiments.

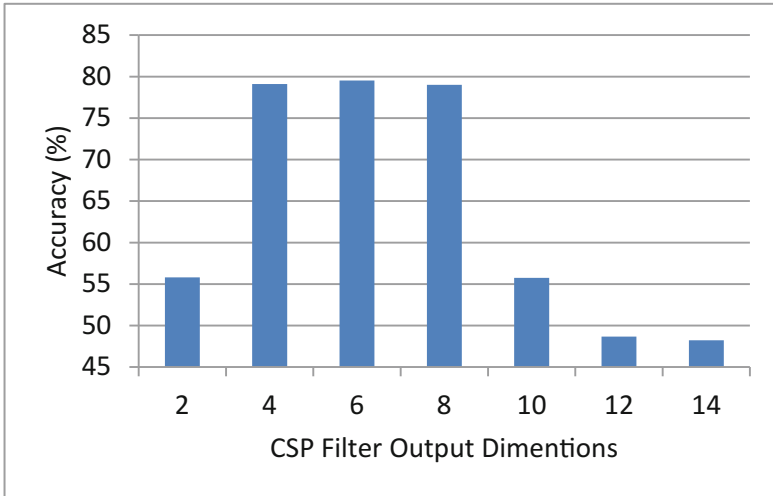


Fig. 6. Mean accuracy (%) for different numbers of output channels in the CSP filter.

4.3 Evaluating How the Classifier Parameters Improves Classification Accuracy

The last analyses were performed on the Nu and gamma parameters which can improve the accuracy of Nu-SVM with RBF kernel [20]. The value of the Nu parameter varies from 0 to 1 [21], and consequently the tests were concentrated in this range. The experimental results using the 24 MI trials dataset are shown in Fig. 7. It was observed that the best mean accuracy value was obtained with Nu equal to 0.5. Next, in order to verify the range near to this value, the analysis was extended, as can be seen in Table 1.

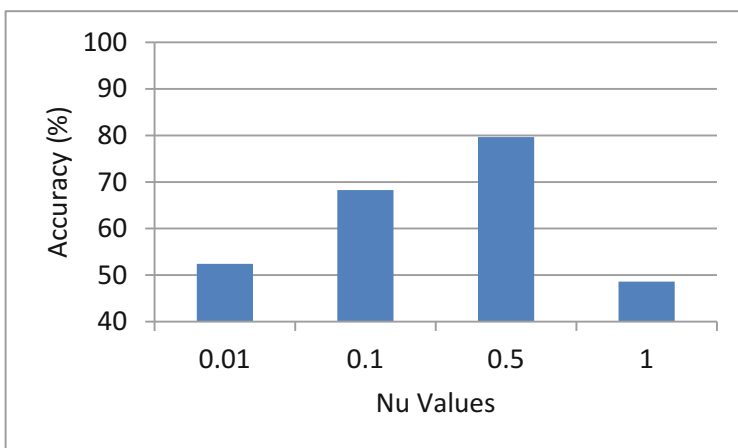


Fig. 7. Mean accuracy (%) for different values of Nu.

Table 1. Mean accuracy (%) and standard deviation to all subjects for values of Nu near to 0.5.

Nu	0.3	0.35	0.4	0.45	0.5	0.6	0.7
Mean	78.451	80.136	80.799	80.647	79.668	75.357	71.819
S. deviation	6.765	7.201	5.874	2.368	2.432	2.068	1.455

To evaluate the gamma values, we maintained the three values of Nu that presented the best accuracy: 0.35, 0.4, and 0.45. Then, a grid-search was performed and various pairs of (Nu, gamma) values were tested, as shown in Fig. 8. In the end, the (0.35, 70) pair was chosen, once it presented the best mean accuracy.

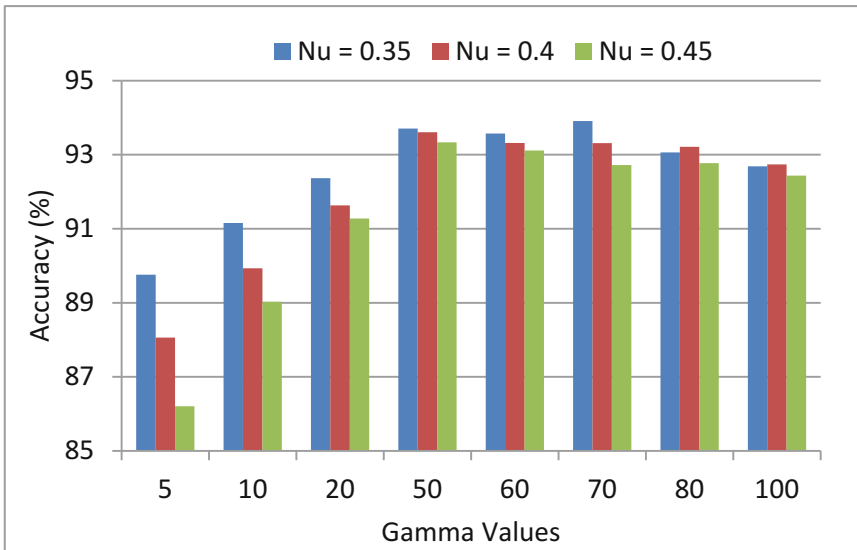


Fig. 8. Mean accuracy (%) for gamma and Nu values.

Finally, the proposed procedure and the OpenVibe MI standard classification scenario were executed for each subject separately. Some characteristics of the OpenVibe MI Standard include a bandpass Butterworth filter with 8–30 Hz frequency band; six output channels in the CSP filter; and the Linear Discriminant Analysis algorithm as the classification approach. The results per subject in the 24 MI trials dataset with a 5-fold cross validation are presented in Fig. 9. The proposed procedure and the OpenVibe framework achieved 93,91% and 64,95% of mean accuracy, respectively.

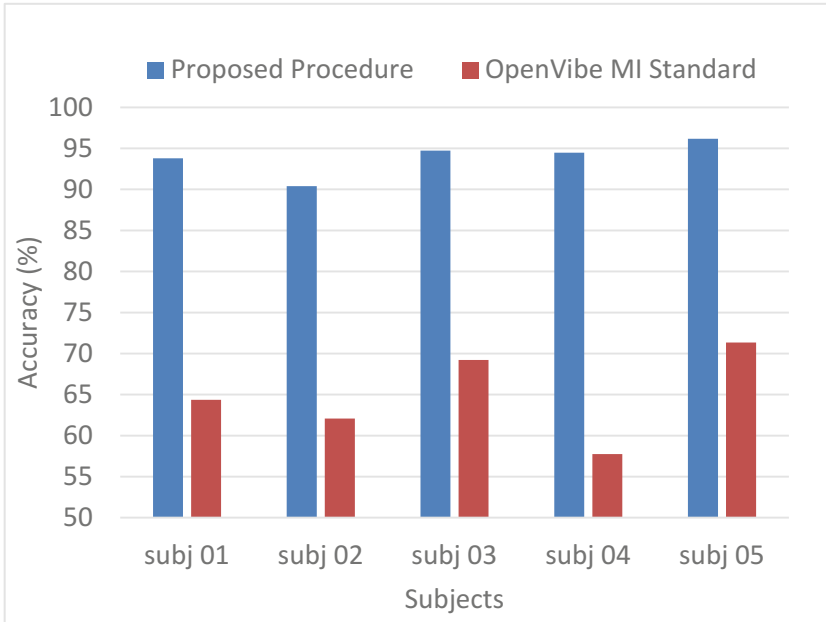


Fig. 9. Mean accuracy (%) results obtained using the proposed procedure and the OpenVibe standard scenario for MI classification.

5 Conclusions and Future Works

This work proposed a discrimination procedure of two EEG signal classes produced by imagined movements through the OpenVibe software. A MI database was built and several characteristics were considered in order to create the proposed procedure. These aspects were carefully analysed, ranging from the signals acquisition to the SVM classifier parameters configuration. The procedure obtained an increase of 28,96% in the mean accuracy, regarding to the OpenVibe MI standard scenario, showing that to improve a BCI system performance these characteristics need to be considered, since they are directly related to the classification results. In the future, we intend to create BCI systems that run automatically the procedure as well as to apply the procedure to other databases. In addition, a framework of the proposed procedure will be created in a way unrelated to OpenVibe and some applications will be implemented to test its online performance.

References

1. Wolpaw, J.R., Birbaumer, N., Heetderks, W.J., McFarland, D.J., Peckham, P.H., Schalk, G., Donchin, E., Quatrano, L.A., Robinson, C.J., Vaughan, T.M.: Brain-computer interface technology: a review of the first international meeting. *IEEE Trans. Rehabil. Eng.* **8**(2), 164–173 (2000)
2. Djemal, R., Bazyed, A.G., Belwafi, K., Gannouni, S., Kaaniche, W.: Three-class EEG-based motor imagery classification using phase-space reconstruction technique. *Brain Sci.* **6**(3), 36 (2016)
3. Abdalsalam, M.E., Yusoff, M.Z., Kamel, N., Malik, A., Meselhy, M.: Mental task motor imagery classifications for noninvasive brain computer interface. In: *Intelligent and Advanced Systems, ICIAS, Kuala Lumpur*, pp. 1–5. IEEE (2014)
4. Ma, Y., Ding, X., She, Q., Luo, Z., Potter, T., Zhang, Y.: Classification of motor imagery EEG signals with support vector machines and particle swarm optimization. *Comput. Math. Methods Med.* **2016**, 8 (2016)
5. Yang, Y., Kyrzyzov, O., Wiart, J., Bloch, I.: Subject-specific channel selection for classification of motor imagery electroencephalographic data. In: *Acoustics, Speech and Signal Processing, ICASSP, Vancouver*, pp. 1277–1280. IEEE (2013)
6. Sivakami, A., Devi, S.S.: Analysis of EEG for motor imagery based classification of hand activities. *Int. J. Biomed. Eng. Sci.* **2**(3), 11–22 (2015)
7. Pfurtscheller, G., Brunner, C., Schlögl, A., Da Silva, F.L.: Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage* **31**(1), 153–159 (2006)
8. Pfurtscheller, G., Da Silva, F.L.: Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.* **110**(11), 1842–1857 (1999)
9. OpenVibe. Homepage. <http://openvibe.inria.fr/>. Accessed 23 Jan 2018
10. Jiralerspong, T., Liu, C., Ishikawa, J.: Identification of three mental states using a motor imagery based brain machine interface. In: *Computational Intelligence in Brain Computer Interfaces, CIBCI, Orlando*, pp. 49–56. IEEE (2015)
11. Hurtado-Rincon, J., Rojas-Jaramillo, S., Ricardo-Cespedes, Y., Alvarez-Meza, A.M., Castellanos-Dominguez, G.: Motor imagery classification using feature relevance analysis: an Emotiv-based BCI system. In: *Image, Signal Processing and Artificial vision, STSIVA, Armenia*, pp. 1–5. IEEE (2014)
12. TCT Webpage. https://www.trans-cranial.com/local/manuals/10_20_pos_man_v1_0_pdf.pdf. Accessed 23 Jan 2018
13. Wolpaw, J.R., Boulay, C.B.: Brain signals for brain–computer interfaces. In: Graimann, B., Pfurtscheller, G., Allison, B. (eds.) *Brain-Computer Interfaces. The Frontiers Collection*, pp. 29–46. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02091-9_2
14. Szachewicz, P.: Classification of motor imagery for braincomputer interfaces. Poznan University of Technology, Institute of Computing Science, Poznań (2013)
15. Herman, P., Prasad, G., McGinnity, T.M., Coyle, D.: Comparative analysis of spectral approaches to feature extraction for EEG-based motor imagery classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **16**(4), 317–326 (2008)
16. Carrera-Leon, O., Ramirez, J.M., Alarcon-Aquino, V., Baker, M., D’Croz-Baron, D., Gomez-Gil, P.: A motor imagery BCI experiment using wavelet analysis and spatial patterns feature extraction. In: *Engineering Applications Workshop, WEA, Bogota*, pp. 1–6. IEEE (2012)

17. Vargic, R., Chlebo, M., Kacur, J.: Human computer interaction using BCI based on sensorimotor rhythm. In: Intelligent Engineering Systems, INES, Bratislava, pp. 91–95. IEEE (2015)
18. Mathur, A., Foody, G.M.: Multiclass and binary SVM classification: implications for training and classification users. *IEEE Geosci. Remote Sens. Lett.* **5**(2), 241–245 (2008)
19. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Muller, K.R.: Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Sig. Process. Mag.* **25**(1), 41–56 (2008)
20. Huang, C.L., Wang, C.J.: A GA-based feature selection and parameters optimization for support vector machines. *Expert Syst. Appl.* **31**(2), 231–240 (2006)
21. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. *Neural Comput.* **12**(5), 1207–1245 (2000)



WebBCI: An Electroencephalography Toolkit Built on Modern Web Technologies

Pierce Stegman^(✉), Chris Crawford, and Jeff Gray

University of Alabama, Tuscaloosa, AL, USA
{pwstegman, crawford, jggray}@ua.edu

Abstract. Recent advances in electroencephalography (EEG) based brain-computer interfaces (BCIs) have led to applications that allow users to control devices such as wheelchairs, prosthetic arms, drones, and gaming systems using cognitive commands. However, software environments used to create these technologies are often designed for expert developers. This research paper investigates the feasibility of JavaScript as a development platform for non-critical BCI systems. We also discuss the current issues with JavaScript-based BCIs and introduce a new library, WebBCI, which is designed to take the initial step towards addressing these issues. Initial benchmarks of WebBCI suggest JavaScript can run common EEG and BCI methods such as band power extraction, common spatial pattern, and linear discriminant analysis in real-time on an array of devices, including mobile phones.

Keywords: Brain-computer interface (BCI) · Electroencephalography (EEG) JavaScript

1 Introduction

Recent advancements in brain-computer interface (BCI) technology have led to the development of many computer systems that respond to brain activity rather than more traditional motor inputs, such as a mouse and keyboard. Such systems have allowed users to control prosthetic arms, fly drones, and even play video games using their thoughts [1–3]. Many different hardware and software solutions exist for the creation of such systems. A popular hardware solution is an electroencephalography (EEG) headset. An EEG headset collects central nervous system activity from electrodes placed on the scalp and reports the results back to a computer which analyzes and respond to the data [4].

There exist many different software solutions to parse and analyze EEG data, each with different goals and target users in mind. The main goals of these systems can be divided into three categories: (1) to aid in the research of data processing methods, (2) to make BCI system design available to a wider audience, and (3) to develop systems that can be conveniently deployed to end-users. However, there exists a lack of software solutions that focus on goal (3). As such, while BCI software platforms support research and development, few systems focus on development for consumer accessible technologies.

This paper presents a web-based JavaScript approach to BCI application development. JavaScript can run on mobile devices, tablets, desktop computers, and servers, as well as directly in a web browser, making software written in JavaScript available across multiple platforms [5–7]. As such, JavaScript serves well for the basis of portable and accessible BCI systems. However, JavaScript’s mathematical processing capabilities are extremely limited, making the implementation of common EEG signal processing methods difficult when compared to alternative tools such as MATLAB.

WebBCI is a JavaScript library that provides the basic tools necessary to run a BCI system entirely within a web browser. WebBCI builds upon existing JavaScript mathematical libraries such as Math.js [8] and Numeric JavaScript [9], adding BCI-specific paradigms such as common spatial pattern (CSP), machine learning tools such as linear discriminant analysis (LDA), and signal processing methods such as power spectral density (PSD) and band power extraction. WebBCI also includes tools to interface with and manipulate data streamed from EEG headsets. It is published as a Node.js module on NPM and can be run within Node.js or used within a web browser [10].

2 Related Work

A popular tool for BCI system design is BCILAB, a MATLAB toolbox developed predominately by Christian Kothe at the Swartz Center for Computational Neuroscience out of the University of California San Diego [11]. It contains numerous EEG signal processing methods and includes a GUI to aid in the design of the BCI system. However, BCILAB is research-oriented, and as such, BCIs developed through BCILAB may be difficult to run outside of the laboratory environment. Additionally, BCI lab requires a user to purchase MATLAB and understand MATLAB’s syntax and usage. This may serve as a barrier to those with limited programming knowledge and to the distribution of BCI systems.

There exist multiple systems that aim to make BCI system design easier for those with limited programming knowledge. OpenViBE runs on a user’s desktop and features a drag-and-drop style interface where users can interconnect different BCI tools and components [12]. This makes OpenViBE ideal for researchers and students who wish to explore BCI design, but lack programming knowledge.

Another system, NeuroBlock [13, 30], is geared towards students and BCI education in the classroom. NeuroBlock is a block-based programming language with an in-browser user interface. It supports simple EEG methods and allows the student to control an on-screen character by designing a program using a drag-and-drop block-based programming language that uses EEG signals as input [13]. However, while OpenViBE and NeuroBlock take great steps towards making BCI research and education available to a wider audience, the BCIs designed with these tools cannot be compiled easily or run outside their original environments. A BCI designed for the browser in JavaScript is a solution to this problem.

Uri Shaked, a Google Developer Expert for Web and Cloud Technologies [14], has published JavaScript tools for interfacing with the popular Muse EEG headband through both Web Bluetooth and Lab Streaming Layer (LSL) [15, 16]. Currently, users have to install desktop software such as Muse Direct, which forwards data to any

applications desiring input from an EEG device [17]. However, the introduction of Web Bluetooth allows for the collection of EEG data within a web environment, removing the necessity of such desktop software.

NeuroJS is a GitHub organization with a focus on “Neuroscience research done with JavaScript” [18]. Their development of a web-based dashboard for visualizing OpenBCI data is a step towards web-based BCIs [19]. The dashboard is built on `dsp.js`, a digital signal processing library, and allows users to view the data streamed from the OpenBCI in both the time and frequency domains. However, while the visualizations are rendered in the web, a Node.js backend is still required to collect the data from the device.

`Math.js` is a mathematical processing library for JavaScript. It supports matrix operations with arbitrary precision and includes many statistical methods, making it ideal for the mathematical basis of a web-based BCI [8]. `Math.js` can also be extended with other libraries such as `Numeric JavaScript`, which includes methods for calculating eigenvalues and eigenvectors [9]. Such methods are necessary for many dimensionality reduction and signal processing techniques needed in BCI applications.

Machine learning is another large part of BCI design. Tools such as `ml.js` implement many machine learning methods in JavaScript, including principal component analysis (PCA), support vector machines (SVMs), K-Nearest Neighbor (KNN), and artificial neural networks (ANNs), among many others [20]. Tools such as `ConvNetJS` also support deep learning within the browser, allowing for complex classification algorithms to be run [21].

Technologies such as Web Bluetooth take a great step towards web-based EEG data collection, and the NeuroJS OpenBCI dashboard demonstrates the basics of signal processing in the browser. Tools such as `Math.js`, `Numeric JavaScript`, and `ml.js` lay the mathematical framework required for a web-based BCI. `WebBCI` builds on these existing tools by implementing the methods necessary for BCI-specific signal processing and by providing functions necessary to interface with and manipulate data from an EEG headset. `WebBCI` also aims to provide a framework for future web-based BCI development.

3 Overview

BCI technology features a layered approach to software development. An EEG-based BCI project starts with the development of hardware to collect central nervous system activity. Examples of such hardware include the Muse and OpenBCI EEG headsets, which collect EEG data from the scalp and forehead [22, 23]. Advancements in these technologies could lead to cleaner data and more insight, furthering the BCI field.

After data has been collected, it must be processed. Tools such as `MATLAB` and `BCILAB` serve well for the researching of such data processing methods. `MATLAB`'s extensive mathematical tools allow for rapid prototyping of new methods and `BCILAB` allows for data processing steps to be chained together easily into a BCI system. Finally, the BCI application must be distributed. One method is to translate the `MATLAB`-based BCI to a compiled language such as `C`, where the system can run across an array of hardware. However, `C` can be difficult to learn and is prone to errors

that can be difficult to solve for the more novice programmer. The goal of WebBCI is to aid in this deployment layer of BCI system design.

WebBCI is built around the following design goals:

1. Portable, modular, and easy to use
2. Open source and extensible
3. Able to interface well with existing JavaScript solutions

For WebBCI to be portable and accessible across an array of hardware devices, JavaScript was chosen as the foundation for WebBCI. JavaScript's cross platform support allows a single application to run on servers, mobile phones, and desktop computers. Additionally, JavaScript is a higher-level language and may be easier to learn than C or C++. Finally, unlike other higher-level languages, such as Python or Java, JavaScript can run across multiple environments without the need for the user to install additional software.

WebBCI contains many of the common mathematical tools required for BCIs, including the Fast Fourier Transform (FFT) for frequency analysis, CSP for signal separation, and LDA for signal classification. These methods build upon existing JavaScript mathematical libraries such as `fft.js`, `Math.js`, and `Numeric JavaScript`. WebBCI's modular design also allows for new JavaScript libraries, such as neural network libraries, to be integrated into the system. These methods reduce the need for a strong mathematical background, and the student or developer can instead focus on the design of their system, requiring only a higher-level understanding of the underlying mathematical concepts. A full list of included methods can be seen in Tables 1, 2 and 3 within Sect. 4.

The source code for WebBCI is open source and available on GitHub [24]. The code is documented and allows for others to contribute to and extend WebBCI with their own methods.

4 System Architecture

WebBCI currently provides methods in three categories: network operations and data acquisition, data storage and manipulation, and mathematical tools for processing and classification.

WebBCI currently supports the open sound control (OSC) protocol for the collection of data from EEG devices. Many popular EEG devices support OSC, including the Muse and OpenBCI headsets [17, 25]. This data is collected into a 2-dimensional array with EEG samples as rows and electrodes as columns. Table 1 enumerates WebBCI's current networking and data acquisition methods.

Table 1. Networking operations of WebBCI.

Method name	Description
<code>oscCollect</code>	Receives data over OSC and returns it as an array
<code>oscHeaderScan</code>	Scans the network for OSC headers and returns a list of found headers
<code>oscStream</code>	Calls specified callback functions when data with specified OSC headers is seen
<code>wait</code>	Places a delay before or after data is collected from the network

A fully functional example of how data can be collected from an EEG headset using WebBCI can be seen in Appendix A.

After data has been collected via the networking tools in WebBCI, it can then be processed using WebBCI's mathematical tools. Table 2 shows the current methods implemented within WebBCI.

Table 2. Mathematical functions of WebBCI.

Method name	Description
cspLearn	Learns a common spatial pattern from EEG data
cspProject	Projects new EEG data using the parameters computed with cspLearn
generateSignal	Generates a signal with specified frequencies and their respective amplitudes
ldaLearn	Returns the result of running linear discriminant analysis on a dataset of feature vectors
ldaProject	Classifies a feature vector using the parameters computed with ldaLearn
features	A namespace for EEG feature extraction methods
psd	Returns a power spectral density (PSD) array for a given signal
psdBandPower	Returns the average power within a frequency band in a PSD array
signalBandPower	Returns the average power within a frequency band in a signal

WebBCI also contains data storage and manipulation methods. MATLAB-like syntax for array subscripting via colon notation is provided to make channel and sample selection easier. Data windowing methods are also provided for offline processing and model training. Additionally, WebBCI contains functions that load and save arrays as CSV files for future use. A complete list of data manipulation functions is given in Table 3.

Table 3. Data manipulation methods of WebBCI.

Method name	Description
subscript	Applies MATLAB-style matrix subscripting with colon notation to an array
windowApply	Divides a 2-dimensional array along rows into windows of a specified size and overlap amount and applies a given function to each window, returning the results for each window as an array
saveCSV	Saves a 2-dimensional array to a CSV file
loadCSV	Loads a CSV file into a 2-dimensional array
round	Returns a new array with all data points rounded to the specified number of decimal places
toFixed	Returns a new array of string representations of values with zeros padding to the specified number of decimal places
toTable	Returns an ASCII table representation of the array

5 System Performance

To ensure the portability of the WebBCI system, we ran performance benchmarks across a variety of devices. We tested three common methods used within BCIs: band power extraction, CSP, and LDA. Additionally, we tested these methods with different data sizes to see how their performance scaled with larger data sets. The devices chosen for the benchmark were a Lenovo IdeaCentre Desktop PC, a Microsoft Surface Book, a Google Nexus 5X Android Phone, and an Apple iPhone 8. Google Chrome was used to run the benchmark on each device. The specifications of each device can be seen in Appendix B.

The average run time for an operation given a set number of samples was calculated by running each test 100 times and computing the arithmetic mean of the run times. Random data was used during each run so results were not affected by caching or branch prediction.

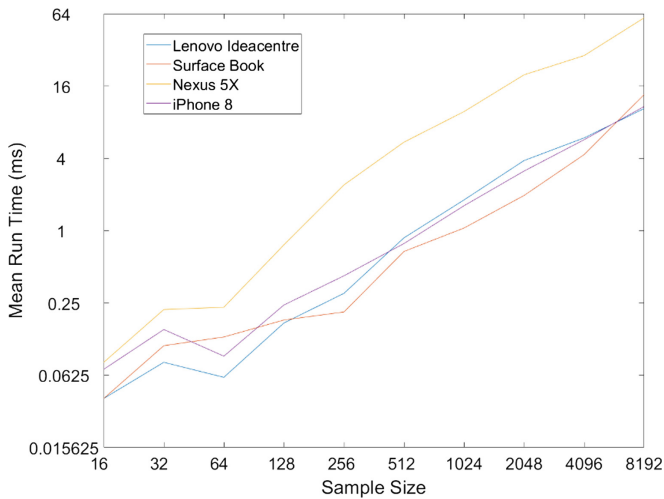


Fig. 1. Average band power extraction benchmark run time given a set number of samples over 16 channels.

Figure 1 shows the processing speed of WebBCI when calculating the average power in a frequency band. A simulated 16 channels of data were used for each test.

For a system to run in real-time, it must process samples faster than it receives them from the EEG headset. Devices such as the Muse and OpenBCI have sample rates

between 220 and 250 Hz [26, 27]. As seen in Fig. 1, every device can perform band power extraction on an equivalent 1 s of data in less than 4 ms.

To test the possibility of applying WebBCI to more complex systems, such as those used for motor imagery, we ran performance benchmarks for both CSP and LDA, as these are the two commonly used functions to create a basic motor imagery BCI [28, 29]. A simulated 16 channels of randomized data were used in this benchmark. The results of this benchmark can be seen in Figs. 2 and 3.

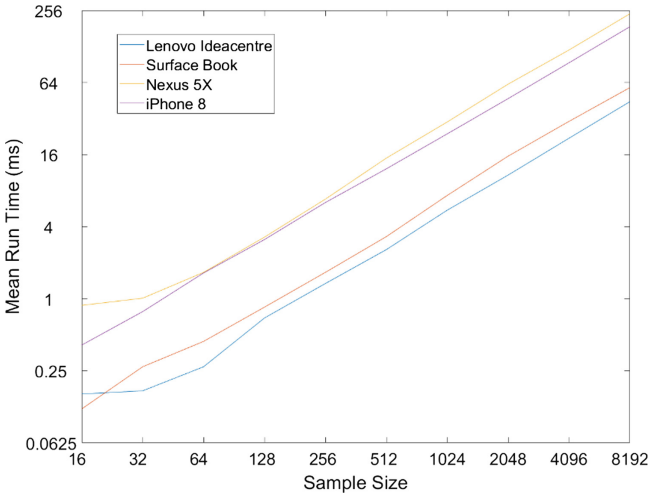


Fig. 2. Average CSP benchmark run time given a set number of samples over 16 channels.

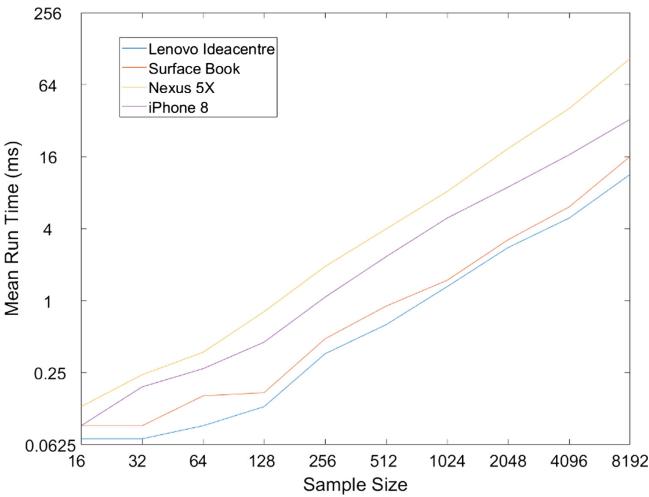


Fig. 3. Average LDA benchmark run time given a set number of samples over 16 channels.

This performance benchmark shows that basic motor imagery BCIs may be possible within a web environment. It also shows the future potential of JavaScript-based mobile BCI applications. Additional testing and research is necessary to test the possibility of potential mobile BCI systems.

6 Discussion and Limitations

WebBCI aims to demonstrate the possibility of a web-based BCI. However, additional BCI-specific functions have yet to be implemented. While WebBCI can be extended with tools such as ml.js for additional classification methods, many signal processing methods must be implemented, including those for filtering and removing artifacts from the data before processing. WebBCI currently serves as a candidate framework for future development and as a demonstration of what may be possible within a web environment.

Additional tests are required to further validate the reported performance benchmarks. Additional browsers, devices, and BCI functions should be evaluated using the presented system. User studies featuring real-time applications developed with WebBCI could also provide further insights regarding usability and performance.

7 Conclusion

While signal processing and BCI methods have been increasingly researched in recent years, there is a gap in the literature for strategies involving web-based BCI systems. JavaScript presents an accessible approach to BCI development and deployment. WebBCI builds on existing JavaScript mathematical libraries to provide a library of functions and a framework for JavaScript-based BCI development. Our performance benchmarks suggest that modern devices and web-browsers may support simple BCI applications.

Appendix A

Below is a functional example of how data can be collected using the WebBCI Node.js module. This example receives OSC data at the specified IP address and port and collects 2500 samples of data containing the header 'Person0/eeg.'

```
var bci = require('webbci');
async function collect() {
  var oscParams=['127.0.0.1', 7000, 'Person0/eeg', 2500];
  console.log('data collection begins in 1 second...');
  await bci.wait(1000);
  var data = await bci.oscCollect(...oscParams);
}
collect().catch(error => console.error(error));
```

After the data has been collected, features can be extracted using the WebBCI data windowing method. An example of this is shown below, where the data is divided into segments of length 32 samples and a step of size 16 samples is taken on every iteration, creating an overlap between windows of 50%. The log of the variance of each channel is used to create a feature vector for each window. These feature vectors are then stored as an array into the variable ‘features.’ Passing ‘false’ as the final parameter specifies that if the array does not evenly divide into windows of size 32, the tail end of the array should be ignored such that it evenly divides into windows.

```
var features = bci.windowApply(samples, 'logvar', 32, 16, false);
```

Appendix B

Table 4. Specifications of devices used to run the performance benchmarks.

Device name	CPU	RAM	Browser	Operating System
Lenovo IdeaCentre	Intel Core i7-6700 CPU @ 3.40 GHz, 4 Cores, 8 Logical Processors	16.00 GB	Google Chrome 64.0.3282.186	Microsoft Windows 10 Home, Version 10.0.15063 Build 15063
Surface Book	Intel Core i7-6600U CPU @ 2.60 GHz, 2 Cores, 4 Logical Processors	8.00 GB	Google Chrome 63.0.3239.132	Microsoft Windows 10 Pro, Version 10.0.16299 Build 16299
Nexus 5X	Qualcomm Snapdragon 808 Processor	2.00 GB	Google Chrome 64.0.3282.137	Android 8.1.0 Build OPM3.171019.014
iPhone 8	A11 Bionic Chip	2.00 GB	Google Chrome 64.0.3282.112	iOS 11.2.5 Build 15D60

References

1. LaFleur, K., Cassady, K., Doud, A., et al.: Quadcopter control in three-dimensional space using a noninvasive motor imagery-based brain–computer interface. *J. Neural Eng.* **10**, 46003 (2013). <https://doi.org/10.1088/1741-2560/10/4/046003>
2. Zander, T.O., Kothe, C.: Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *J. Neural Eng.* **8**, 25005 (2011). <https://doi.org/10.1088/1741-2560/8/2/025005>
3. Müller-Putz, G.R., Pfurtscheller, G.: Control of an electrical prosthesis with an SSVEP-based BCI. *IEEE Trans. Biomed. Eng.* **55**, 361–364 (2008). <https://doi.org/10.1109/TBME.2007.897815>

4. Wolpaw, J., Wolpaw, E.W.: Brain-Computer Interfaces: Principles and Practice. OUP, USA (2012)
5. Tilkov, S., Vinoski, S.: Node.js: using JavaScript to build high-performance network programs. *IEEE Internet Comput.* **14**, 80–83 (2010). <https://doi.org/10.1109/MIC.2010.145>
6. Ghatol, R., Patel, Y.: Beginning PhoneGap: Mobile Web Framework for JavaScript and HTML5. Apress, USA (2012)
7. Charland, A., Leroux, B.: Mobile application development: web vs. native. *Commun. ACM* **54**, 1–5 (2011). <https://doi.org/10.1145/1941487>
8. math.js. <http://mathjs.org/>. Accessed 4 Feb 2018
9. Loisel, S.: Numeric Javascript. <http://www.numericjs.com>. Accessed 4 Feb 2018
10. Stegman, P.: WebBCI. <https://www.npmjs.com/package/webbci>. Accessed 4 Feb 2018
11. Kothe, C.A., Makeig, S.: BCILAB: a platform for brain–computer interface development. *J. Neural Eng.* **10**, 56014 (2013). <https://doi.org/10.1088/1741-2560/10/5/056014>
12. Renard, Y., Lotte, F., Gibert, G., et al.: OpenViBE: an open-source software platform to design, test, and use brain-computer interfaces in real and virtual environments. *Presence Teleoperators Virtual Environ.* **19**, 35–53 (2010). <https://doi.org/10.1162/pres.19.1.35>
13. Crawford, C.S., Gilbert, J.E.: NeuroBlock: a block-based programming approach to neurofeedback application development. In: 2017 IEEE Symposium Visual Languages Human-Centric Computing, pp. 303–307 (2017). <https://doi.org/10.1109/vlhcc.2017.8103483>
14. Uri Shaked. <https://medium.com/@urish/reactive-brain-waves-af07864bb7d4>. Accessed 2 Mar 2018
15. Shaked U muse-js. <https://github.com/urish/muse-js>. Accessed 4 Feb 2018
16. Shaked U muse-lsl. <https://github.com/urish/muse-lsl>. Accessed 4 Feb 2018
17. Muse Direct. <http://www.choosemuse.com/developer/#direct>. Accessed 4 Feb 2018
18. Keller, A., Castillo, A., Kan, J., Shoecraft, A.: NeuroJS. <https://github.com/NeuroJS>. Accessed 11 Oct 2017
19. OpenBCI Dashboard. <https://github.com/NeuroJS/openbci-dashboard>. Accessed 4 Feb 2018
20. ml.js. <https://github.com/mljs/ml>. Accessed 4 Feb 2018
21. Karpathy, A.: ConvNetJS. <https://cs.stanford.edu/people/karpathy/convnetjs/>. Accessed 4 Feb 2018
22. Muse: the brain sensing headband. <http://www.choosemuse.com/>. Accessed 5 Feb 2018
23. OpenBCI - Open Source Biosensing Tools (EEG, EMG, EKG, and more). <http://openbci.com/>. Accessed 5 Feb 2018
24. Stegman, P.: WebBCI. <https://github.com/pwstegman/WebBCI>. Accessed 4 Feb 2018
25. OpenBCI The OpenBCI GUI. http://docs.openbci.com/OpenBCI_Software/01-OpenBCI_GUI. Accessed 11 Oct 2017
26. MuseIO | Available Data. <http://developer.choosemuse.com/tools/available-data>. Accessed 19 Feb 2018
27. Cyton Data Format. http://docs.openbci.com/Hardware/03-Cyton_Data_Format. Accessed 19 Feb 2018
28. Ramoser, H., Müller-Gerking, J., Pfurtscheller, G.: Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehabil. Eng.* **8**, 441–446 (2000). <https://doi.org/10.1109/86.895946>
29. Kothe, C.: Lecture 7.3 Common Spatial Patterns (2013). <https://www.youtube.com/watch?v=zsOULC16USU>. Accessed 5 Feb 2018
30. Crawford, C., Gardner-McCune, C., Gilbert, J.: Brain-computer interface for novice programmers. In: ACM Technical Symposium on Computer Science Education, pp. 32–37. ACM (2018)



A Cross-Brain Interaction Platform Based on Neurofeedback Using Electroencephalogram

Rongrong Zhang and Xiaojie Zhao^(✉)

College of Information Science and Technology,
Beijing Normal University, Beijing, China
zhaox86@163.com

Abstract. Cross-brain neural synchronization has widely been found between participants during social interactions and is suggested to play an important role in human social interactions. Neurofeedback technology feeds the neural signatures of a participant back to himself to modulate his own brain activity. Researches have applied the technology into cross-brain interactions using functional near-infrared spectroscopy (fNIRS) and let two participants do collaborative tasks using brain activities. However, there are few studies in terms of cross-brain interaction based on Electroencephalogram (EEG) signals using neurofeedback technology. In this study, we developed a cross-brain interaction platform based on EEG signals using neurofeedback technology. The platform allows the participants to achieve cross-brain interaction directly with the medium of neurofeedback instead of other participants' body languages or sounds. It was validated with an experiment using a "tug-of-war" game. Through the offline analysis, synchronization between the subjects were found at beta frequency bands across the brains. Cross-brain synchronization reflects the interaction state across the brains and may reflect the strategy that the participants choose. This study is still a preliminary work and needs further work to do.

Keywords: Cross-brain interaction · EEG · Neurofeedback

1 Introduction

Social interaction plays a fundamental role in our daily lives. Simultaneously measuring multiple brains, cross-brain neural synchronization has been found between subjects during various social interactions [1, 18, 19] such as face-to-face communications [4] and musical improvisation on the guitar [5]. These findings suggest that cross-brain synchronization may play an important role in human social interaction. For groups with social barriers, cross-brain interactions in social interactions are abnormal [6]. The study of cross-brain interaction plays an important role to reveal the neural mechanism of human social interaction, and may be helpful to find a new therapy for social disorder such as autism.

According to the existing research, there are mainly two approaches for cross-brain interactions. The first typical approach is to communicate directly using brain signal. Rajesh has applied EEG to send information from one subject, and used transcranial

magnetic stimulation (TMS) to let another subject receive the message [7]. Another approach for cross-brain interaction is to rely on the third-party tools such as collaborate BCI and feedback. Wang has proposed a collaborative paradigm to improve overall BCI performance to response to the direction cue as fast as possible by integrating information from multiple users [8]. Duan has built an experimental platform on the basis of functional near infrared spectroscopy (fNIRS) which allows the two subjects to interact with each other through competing or collaborative tasks such as the tug-of-war game [3].

Neurofeedback is an approach to investigate the relationship between brain activity and behavior, which extracts the neural characteristics of the brain signals and feeds back them to the subjects in visual or auditory ways [2]. It feeds back the neural signatures of a participant to allow him to modulate his own brain activity. Unlike traditional brain imaging studies which uses the “behavioral manipulation – brain observation” paradigms, neurofeedback enables researchers to manipulate the brain activity as an independent variable, which can provide more causal insights into the relationship between brain and behavior. It has been shown that there is converging evidence that a single participant’s brain activity can be self-regulated with neurofeedback technology [9, 10]. In 2004, Goebel for the first time extend neurofeedback from single-person context to multi-person situation and this pioneering work allows multiple subjects simultaneously self-regulate their own neural activities in a social interaction environment [9, 10]. Duan has gone a step further and applied the neurofeedback to two subjects using fNIRS aiming to explore the relationship between the cross-brain neural synchronization and social behavior [3].

Cross-brain synchronization has widely been found during interactive activities. It has been used to provide evidence for the involvement of the brain regions across the subjects for processing the information during the interaction and reveal the neural mechanism behind the social contact. To analyze the synchronization, methods such as phase locking value (PLV), wavelet coherence are widely used. Because human brain is a nonlinear, chaotic and nonstationary system, phase locking is an appropriate approach to quantifying interactions. The most commonly used phase interaction measure is the phase locking value which has been used to measure brain synchronization [22]. Using PLV to measure synchronization, Szymanski has reported increased inter-brain phase synchronization in joint attention relative to individual attention during a visual search task and interpret the findings as neural substrates of social facilitation [21]. Jarhwan has used PLV to analyze the synchronization across the brains and reveals that the right temporal-parietal cortical region, might play an important role in the social interactions of autism spectrum disorder patients [11]. Joy Hirsch has used wavelet analysis and cross-brain coherence analysis to explore the mechanism behind eye-to-eye contact [12]. They found the cross-brain coherence increased for signals originating within left superior temporal, middle temporal and supplementary motor cortices of both interacting brains. The findings reveal a network that mediates neural responses during eye-to-eye contact between dyads.

In this study, a cross-brain interaction platform was constructed based on EEG and neurofeedback technology. The platform allows the participants to achieve cross-brain interaction directly with the medium of neurofeedback instead of other participants’ body languages or sounds. To test the platform, we conducted an experiment using a

“tug-of-war” game. Wavelet coherence analysis was applied to analyze the synchronization between the specific electrodes across the brains, and reveals the synchronization between the subjects during the game.

2 Materials and Method

2.1 Cross-Brain Interaction Platform

The whole platform of the cross-brain interaction based on EEG is shown in Fig. 1. We used 32 channels g.Nautilus wireless EEG acquisition equipment (g.tec medical engineering GmbH, Austria) to simultaneously measure multiple brains’ signals. The raw EEG signals are recorded by the scalps and transmitted to PC stations using Bluetooth protocol. The signals will be merged into a core PC computer in real time using UDP protocol, which are prepared for the subsequent calculation including online analysis module and feedback module. An online analysis module is used to extract and calculate the feedback signal based on EEG (Fig. 4A). The module is developed using SDK provided by g.tec company with MATLAB (R2014a, The MathWorks Corporation). The feedback module is used to give the subjects a more directly and intuitive form of the current interactive situation across the brains to allow the subjects to do self-regulating and is built using Java swing framework (Java version “1.8.0_144”, Java Runtime Environment build 1.8.0_144-b01).

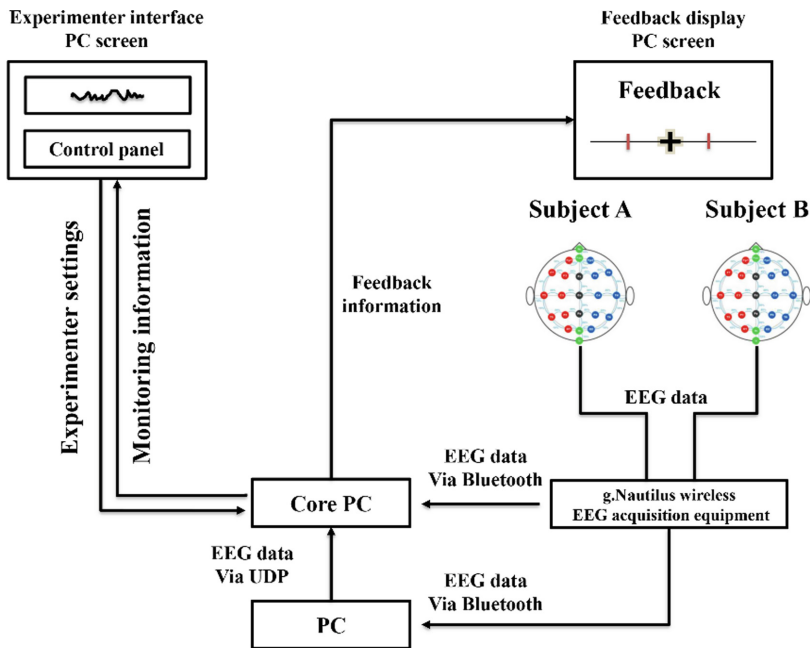


Fig. 1. The EEG-based interaction platform cross two subjects in the hardware level

2.2 The Experiment Design

To test the platform, we conducted an experiment using the “tug-of-war” game which let the two subjects fought a cross-brain “tug-of-war” against each other [3]. An overview of the experimental scenario is shown in the Fig. 2B. Two male right-handed volunteers (age 24 years) without brain diseases or bad habits are recruited from Beijing Normal University participant in this experiment, and gave their written informed consent prior to their inclusion in the study. Two subjects were required to engage through motor imagination in the event of physical inactivity. A line with a cursor was displayed on the screen as is shown in Fig. 2A. Initially, the cursor was positioned at the midpoint of the line. Each subject was allowed to imagine only left or right and use a strategy to pull the cursor on the screen to their side. The difference between the amplitudes of their brain activities at each time point corresponded to the amount of the cursor’s shift. The entire experiment consists of a preparation process lasting 30 s, and 5 rounds to conduct the game. The preparation process was used for the subjects to adjust the state of their minds while facilitating the follow-up baseline correction. Each round consists of a resting and a tasking phase, where the duration of the rest period is 40 s and the duration of the task is also 40 s.

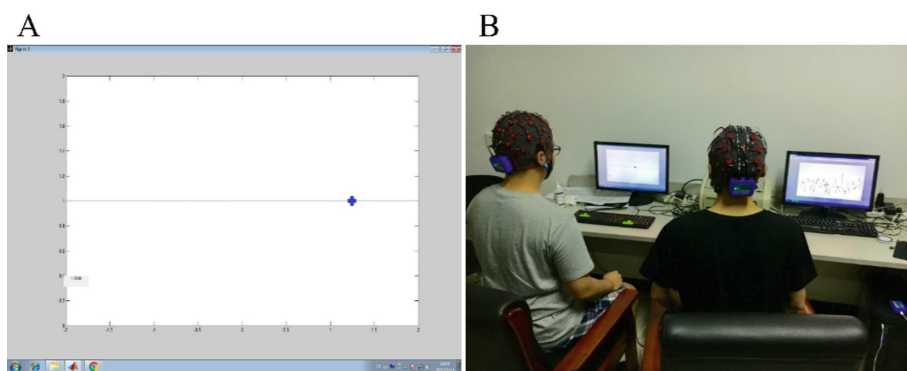


Fig. 2. The feedback screen for participants (A) and an overview of the experimental scenario (B).

2.3 Online Analysis

The online data processing was programmed with MATLAB and performed on a PC station (Microsoft Windows 7 operating system). The data was filtered by a FIR band-pass filter at 0.01–40 Hz and baseline corrected. Based on event related synchronization and desynchronization principle, for one subject, the difference of band power between the C3, C4 at beta band was calculated as BP_A for subject A and BP_B for subject B in real time to determine directions the subject was thinking [13, 14, 17]. The difference between the calculation of bandpower between subjects, which is $BP_A - BP_B$, was used as the feedback information to shift the cursor on the screen. The feedback calculation was programmed as the MATLAB m-files to embed into modules. The experimenter could monitor the subjects’ brain activity online which

is represented as the trajectory of cursor, as shown in Fig. 1. The trajectory of the cursor was recorded for subsequent analysis. Because of the limitation of performance in the simulation system, the feedback module is built independently which is connected to the feedback calculation module using UDP communication protocol.

2.4 Offline Analysis

In order to explore the neural synchronization between participants during the interactions, an offline analysis is performed with a wavelet coherence analysis [12, 15, 20]. Wavelet analysis decomposes a time varying signal into frequency components. Cross brain coherence is measured as a correlation between two corresponding frequency components, and is represented as a function of the period of the frequency components. It is a great tool to track the correlation between the band pass filtered components of time series signals and the equations are as follows:

$$WC(t, f) = \frac{|SW_{XY}(t, f)|}{\sqrt{|SW_{XX}(t, f)| |SW_{YY}(t, f)|}} \quad (1)$$

Where

$$SW_{XY}(t, f) = \int_{t-\frac{\delta}{2}}^{t+\frac{\delta}{2}} W_X(\tau, f) W_Y^*(\tau, f) d\tau \quad (2)$$

$$SW_{XX}(t, f) = \int_{t-\frac{\delta}{2}}^{t+\frac{\delta}{2}} W_X(\tau, f) W_X^*(\tau, f) d\tau \quad (3)$$

$$SW_{YY}(t, f) = \int_{t-\frac{\delta}{2}}^{t+\frac{\delta}{2}} W_Y(\tau, f) W_Y^*(\tau, f) d\tau \quad (4)$$

And $W_X(\tau, f)$ is the complex Morlet wavelet transform, the definition is as follows:

$$W_X(\tau, f) = \int x(u) \varphi_{\tau, f}^*(u) du \quad (5)$$

Where

$$\varphi_{\tau, f}(u) = \sqrt{f} \cdot e^{j2\pi f(u-\tau)} \cdot e^{-\frac{(u-\tau)^2}{\delta^2}} \quad (6)$$

and $W_X^*(\tau, f)$ is the conjugate of the $W_X(\tau, f)$.

4-layer wavelet packet decomposition is performed with db4 wavelet base to extract beta rhythm waves from C3 and C4 electrodes for all participants [16]. As shown in Fig. 3, the cross coherence was computed across the C3, C4 channel on the same side and different side between the subjects. Namely the coherence was computed between C3_A – C3_B, C4_A – C4_B, C3_A – C4_B, C4_A – C3_B.

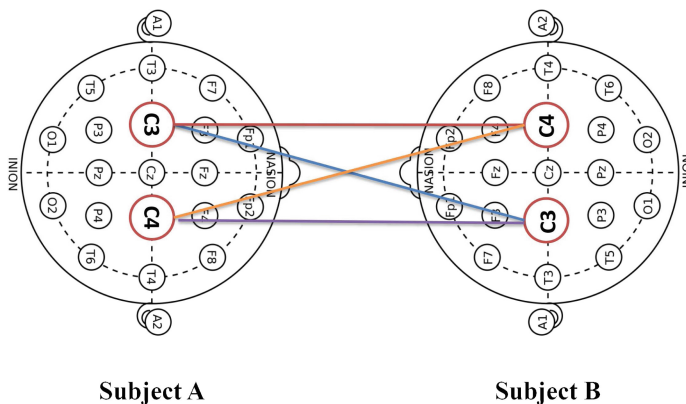


Fig. 3. The coherence computation between subject A and B in offline analysis.

3 Result

All the participants reported that playing the tug-of-war game with another person was very interesting and they can perform the motor imagery well. They reported that when their attentions were highly concentrated, they can pull the cursor back to their side easily. On the contrary, if they lost their attention, the cursor would be pulled to the opposite side.

As shown in Fig. 4B, the top subgraph indicates the track of cursor’s movement controlled by subjects through online analysis. It expands the trajectory of the

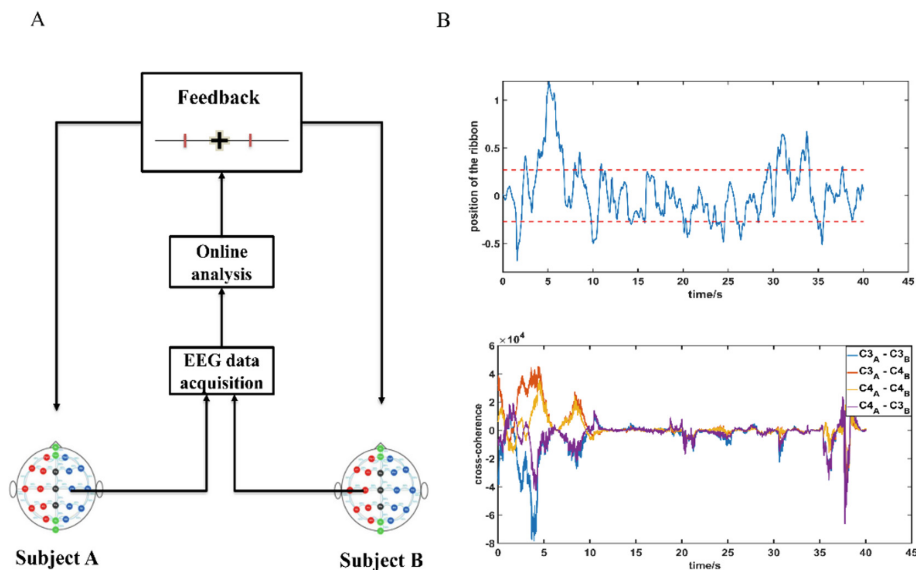


Fig. 4. The structure of the cross-brain interaction platform (A) and the results of online and offline analysis in the tug-of-war experiment (B). (Color figure online)

movement to the time dimension and represents that as time goes by, the cursor moves left to right. The blue line represents the movement of the cursor. If the cursor has been pulled back over three-fifths of the half-length of the line by one subject, we determine him the winner at that time [3]. The red dash lines represent the boundaries. Outside the boundaries of the red lines indicates the winner while inside the boundaries of the red lines cannot indicate who is winner. It can be found that the blue line is mainly outside the boundaries during 0 s–10 s and 28–31 s. The down subgraph represents the cross-coherence between specific channels across the subjects by offline analysis.

Through offline analysis, it can be found that the coherence increased between the electrodes across the brains, such as C3 from subject A and C4 from subject B during 0 s–10 s, 18–25 s and 28–31 s. However, the coherence's polarity of the same side electrodes is opposite to that of the different side. For example, the polarity of the coherence between C3_A – C3_B and C4_A – C3_B is opposite.

4 Discussion

We successfully build a cross-brain interaction platform based on neurofeedback and conduct a tug-of-war game to test the platform. Through the platform, the neural activity of the two subjects was observed and calculated online based on EEG. According to the trajectory recorded during the online game, most of the time two subjects fought against each other equivalently which indicates that the amplitudes of the brain activities across the subjects are roughly similar. Through the questionnaire, we found that when the subjects were distracted, the cursor would move to the other side while the subjects were concentrated, the cursor would move to their side. The subjects expressed that the game was very interesting and novel and they could interact with each other simply using EEG signals through the imagination of left or right to control the cursor to move.

Through the offline analysis, we found that there exists a relationship between the synchronous state of the brains and the movement of the cursor during the game. The higher the synchrony of the beta band of the C3, C4 electrodes' signals across the brains, the better the results to distinguish between the subjects. During the period between 0 s–10 s, we can easily determine the winner in Fig. 4A while the coherence between the two subjects is relatively high. We can't determine the winner during the 10 s–20 s while the coherence between the two is relatively low. In Duan's research, the Pearson correlation coefficient (r) was used to measure the cross-brain relationship between the two participants and they found the neural synchronization during the game [3]. In this study, the synchronization between specific electrodes at beta bands demonstrates that EEG based neurofeedback can be used to explore the cross-brain interaction.

According to the existing research, Jaehwan has found the cross-brain synchronization when the subjects tend to choose different strategies such as cooperation or defection strategy in the prisoner's dilemma game experiment [11]. The PLV between FZ-P8, FCZ-P8, C3-P6 across the brains is higher than that when subjects choose the defection strategy compared with cooperation strategy and the right temporal-parietal cortical region may play an important role in the social interactions of autism spectrum

disorder patients. In this study, the synchronization across the brains may reflect the strategies which subjects have chosen to compete with each other during the game. Because C3, C4 electrodes are relevant to motor imagery [17], we only choose these two electrodes to conduct the experiment and analysis for simplicity and convenience at the initial stage of the study. In the following phases of the study, individual channels should be grouped into anatomical regions based on shared anatomy, which is served to optimize signal-to-noise ratios. Multiple channels electrodes should be taken into consideration of cross-brain coherence analysis not just the specific ones.

5 Conclusions

In summary, the study established a cross-brain interaction platform based on neurofeedback using EEG signals. A validation experiment with a tug-of-war game has demonstrated that the platform can record and integrate two participants' EEG signals to calculate and feedback the cross-brain neural interaction. Cross-brain synchronization has been found during the offline analysis which reflects the interaction state across the brains and may reflect the strategy that the participants choose. This study is still a preliminary work and needs further work to do. On the basis of the present work, next step we plan to investigate the neural features in the cross-brain synchronization which reflects the cooperation or competition strategies.

References

1. Cui, X., Bryant, D.M., Reiss, A.L.: NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation. *Neuroimage* **59**(3), 2430–2437 (2012)
2. Lubar, J.F., Swartwood, M.O., Swartwood, J.N., O'Donnell, P.H.: Evaluation of the effectiveness of EEG neurofeedback training for ADHD in a clinical setting as measured by changes in TOVA scores, behavioral ratings, and WISC-R performance. *Appl. Psychophysiol. Biofeedback* **20**(1), 83–99 (1995)
3. Duan, L., Liu, W.J., Dai, R.N., Li, R., Lu, C.M., Huang, Y.X., Zhu, C.Z.: Cross-brain neurofeedback: scientific concept and experimental platform. *PLoS ONE* **8**(5), e64590 (2013)
4. Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., Lu, C.: Neural synchronization during face-to-face communication. *J. Neurosci.* **32**(45), 16064–16069 (2012)
5. Müller, V., Sängler, J., Lindenberger, U.: Intra- and inter-brain synchronization during musical improvisation on the guitar. *PLoS ONE* **8**(9), e73852 (2013)
6. Tanabe, H.C., Kosaka, H., Saito, D.N., Koike, T., Hayashi, M.J., Izuma, K., Okazawa, H., et al.: Hard to “tune in”: neural mechanisms of live face-to-face interaction with high-functioning autistic spectrum disorder. *Frontiers in human neuroscience* **6**, 268 (2012)
7. Rao, R.P., Stocco, A., Bryan, M., Sarma, D., Youngquist, T.M., Wu, J., Prat, C.S.: A direct brain-to-brain interface in humans. *PLoS ONE* **9**(11), e111332 (2014)
8. Wang, Y., Jung, T.P.: A collaborative brain-computer interface for improving human performance. *PLoS ONE* **6**(5), e20422 (2011)

9. Goebel, R., Sorger, B., Birbaumer, N., Weiskopf, N.: Learning to play BOLD brain pong: from individual neurofeedback training to brain–brain interactions. Organization for Human Brain Mapping, Canada, ON (2005)
10. Goebel, R., Sorger, B., Kaiser, J., Birbaumer, N., Weiskopf, N.: BOLD brain pong: self-regulation of local brain activity during synchronously scanned, interacting subjects. In: 34th Annual Meeting of the Society for Neuroscience (2004)
11. Jahng, J., Kralik, J.D., Hwang, D.U., Jeong, J.: Neural dynamics of two players when using nonverbal cues to gauge intentions to cooperate during the Prisoner’s Dilemma Game. *NeuroImage* **157**, 263–274 (2017)
12. Hirsch, J., Zhang, X., Noah, J.A., Ono, Y.: Frontal temporal and parietal systems synchronize within and across brains during live eye-to-eye contact. *NeuroImage* **157**, 314–330 (2017)
13. Pfurtscheller, G., Da Silva, F.L.: Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.* **110**(11), 1842–1857 (1999)
14. Babiloni, F., Cincotti, F., Marciani, M., Salinari, S., Astolfi, L., Tocci, A., Mattia, D., et al.: The estimation of cortical activity for brain-computer interface: applications in a domestic context. *Comput. Intell. Neurosci.* **2007**, 11 (2007)
15. Lachaux, J.P., Lutz, A., Rudrauf, D., Cosmelli, D., Le Van Quyen, M., Martinerie, J., Varela, F.: Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. *Neurophysiol. Clin./Clin. Neurophysiol.* **32**(3), 157–174 (2002)
16. Ting, W., Guo-zheng, Y., Bang-hua, Y., Hong, S.: EEG feature extraction based on wavelet packet decomposition for brain computer interface. *Measurement* **41**(6), 618–625 (2008)
17. Pfurtscheller, G., Neuper, C.: Motor imagery and direct brain-computer communication. *Proc. IEEE* **89**(7), 1123–1134 (2001)
18. Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., Garnero, L.: Inter-brain synchronization during social interaction. *PLoS ONE* **5**(8), e12166 (2010)
19. Lindenberger, U., Li, S.C., Gruber, W., Müller, V.: Brains swinging in concert: cortical phase synchronization while playing guitar. *BMC Neurosci.* **10**(1), 22 (2009)
20. Burgess, A.P.: On the interpretation of synchronization in EEG hyperscanning studies: a cautionary note. *Front. Hum. Neurosci.* **7**, 881 (2013)
21. Szymanski, C., Pesquita, A., Brennan, A.A., Perdakis, D., Enns, J.T., Brick, T.R., Lindenberger, U., et al.: Teams on the same wavelength perform better: inter-brain phase synchronization constitutes a neural substrate for social facilitation. *Neuroimage* **152**, 425–436 (2017)
22. Aydore, S., Pantazis, D., Leahy, R.M.: A note on the phase locking value and its properties. *Neuroimage* **74**, 231–244 (2013)



Single-Channel EEG Sleep Stage Classification Based on K-SVD Algorithm

Shigang Zuo and Xiaojie Zhao^(✉)

College of Information Science and Technology,
Beijing Normal University, Beijing, China
zhaox86@163.com

Abstract. Sleep stage classification based on visual inspection is non-automatic and subjective resulting in automatic sleep staging by computer is essential for sleep assessment. Especially, single-channel electroencephalogram (EEG) sleep staging has the particular advantage in wearable devices. Sparse representation classification (SRC) can achieve the classification with a liner combination of atoms in an over-complete dictionary and has been widely applied to pattern recognition. An important step of SRC is dictionary training that commonly used K-SVD algorithm has not been used in sleep EEG studies. In this study we introduce K-SVD dictionary training method based SRC into single-channel EEG sleep stage classification and compare the classification performance between the Pz-Oz channel and the Fpz-Cz channel. The results showed that K-SVD based SRC obtained 96.52%, 88.63%, 85.11%, 82.74% and 80.17% classification overall accuracy for 2-6 sleep stages. The assessment results showed that SRC got good performance in EEG sleep staging and Pz-Oz channel performed better than Fpz-Cz channel. Such method is beneficial to the research of sleep monitoring equipment and the study of sleep-related diseases.

Keywords: Sleep stage · Single-Channel EEG
Sparse representation classification · Dictionary training

1 Introduction

Sleep stage classification is of great importance for sleep quality assessment, the diagnosis and treatment of sleep disorder [1]. The manual sleep staging is performed by experts based on polysomnography (PSG) that requires a combination of electroencephalogram (EEG), electrooculogram (EOG), electrocardiogram (ECG), electromyogram (EMG) and other signals. This process is non-automatic and has strong subjectivity, which leads to automatic objectively sleep staging by computer. According to the widely used R&K criteria [2] and AASM criteria [3] about sleep stages, previous studies have combined EEG, EMG, EOG signals from PSG to classify 6 stages (Wake, REM (rapid eye movement), S1–S4) or 5 stages (Wake, REM, N1–N3) [1, 4]. In these neurophysiological signals, EEG is considered to be a more effective assessment signal [5]. However, due to the high cost and poor portability, the multi-channel sleep monitoring device is difficult to be popularized, thus making the sleep stage classification based on single channel EEG very important.

Previous studies have shown that single-channel EEG used for sleep staging included Pz-Oz channel [5, 6], Fpz-Cz channel [7, 8], and both these two channels together [9]. Some researchers indicated that Pz-Oz can be used instead of Fpz-Cz to get better classification results [6] and some concluded that Fpz-Cz channel is the better one [8]. It is unclear yet which channel is the optimal channel for sleep stage classification. Using single channel EEG signal, these studies adopted SVM, random forest, and Adaboost classifiers [6, 7, 9]. Different from these classifiers, sparse representation classification (SRC) can use the training data set to represent the test samples to achieve classification [10], which has been successfully applied on the study of EEG signal. Yu used SRC to detect vigilance in the normal EEG signals, and the classification accuracy was 94.22% [11]. SRC was also used for detection of abnormal EEG [12] and brain-computer interface applications [10, 13]. Additionally, Liu used sparse representation and collaborative representation to extracted features and compared the sleep stage classification performance with 78-dimensional features from two channel EEG signals and got 80.47% accuracy [14]. However, SRC method is still rarely used as a classification method in sleep staging research.

An important step of SRC is dictionary learning, in which the most commonly used method is K-SVD (K Singular Value Decomposition) algorithm. K-SVD algorithm is an iterative algorithm for dictionary atoms updating process based on sparse coding and current dictionary, proposed by Aharen [15]. Liu used the K-SVD algorithm to construct a complete dictionary to distinguish between different brain tasks in the activated brain sources and achieved a good result [16]. Previous study demonstrated that the K-SVD algorithm has a good classification performance on neurophysiological signals [17], but such algorithm has not been reported to be used in sleep stage classification.

In this study, we introduce SRC into single-channel EEG sleep stage classification and compare the performance of Pz-Oz and Fpz-Cz channels with a few features. First, the sample entropy of EEG signal and the variance and kurtosis of EEG rhythm for each epoch are extracted as features. Then the K-SVD algorithm is used to train the dictionary of each single sleep stage, whose size is different according to the number of one stage epochs. According to the reconstruction residual of the coding coefficients, classification accuracy is achieved and the classification performance of different channels is compared.

2 Materials and Method

2.1 Data Description

A publicly available dataset Sleep-EDF downloaded from PhysioNet website [18] is used in this study. The EEG records used in our study include four healthy subjects aged from 21 to 35 years. Each subject includes two EEG channels of Fpz-Cz and Pz-Oz. The EEG data recorded from 10:00 pm to 7:00 am of the next day is used. The EEG signal was divided into segments in 30 s with sample rate of 100 Hz, called epochs. The original sleep stages of these epochs were labeled as AWA (wake stage), S1, S2, S3, S4, REM (rapid eye movement) MVT (movement time), and UNS (unknown state).

2.2 Feature Extraction

For each channel, all EEG data are filtered by a FIR band-pass filter at 0.4–35 Hz. MVT epochs and UNS epochs are deleted directly because the number of them is very small and 4318 epochs are used finally.

For each epoch, 5-layer wavelet packet decomposition is performed with db4 wavelet base to acquire EEG rhythm wave (Table 1). Then we calculate the variance and kurtosis of the rhythm wave, kurtosis can detect a sharp rise or fall in the part of the rhythm. Then the sample entropy of each epoch is also calculated as one feature. Finally these 13 features are normalized to [0, 1] (Fig. 1).

Table 1. Frequency ranges in 5-layer wavelet packet decomposition.

Coefficient set	Frequency range	Related rhythm
D1	25–35	Low-gamma
D2	12.5–25	Beta
D3	6.25–12.5	Alpha
D4	3.125–6.25	Theta
D5	1.5625–3.125	Delta
D6	0–1.5625	Delta

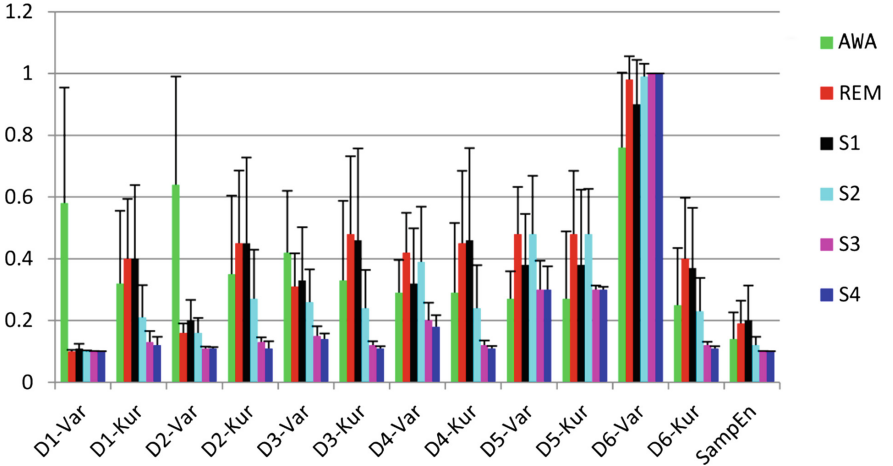


Fig. 1. Normalized feature value. D1–D6 represents the rhythm wave. Var represents variance, and Kur represents kurtosis. SampEn represents sample entropy.

2.3 Dictionary Training Using K-SVD

The procedure of SRC contains two steps: dictionary training and coding classification. For a given training set $\mathbf{Y}_{\text{train}}$, it can be represented by a dictionary \mathbf{D} contains all the information and a corresponding sparse coefficient matrix \mathbf{X} , as Eq. (1).

$$\mathbf{Y}_{train} = \mathbf{D}\mathbf{X} \tag{1}$$

The usage of K-SVD algorithm for dictionary training consists of two parts: sparse coding and dictionary updates. A detailed description of the K-SVD algorithm is shown as the following steps.

- (a) Initialization: initial dictionary $\mathbf{D}^{(0)} \in \mathbb{R}^{m \times K}$, training set $\mathbf{Y} \in \mathbb{R}^{m \times N}$. Let $i = 1$, $\mathbf{X} = 0$, given the upper limit of iteration I.
- (b) Sparse coding: the sparse matrix $\mathbf{X}^{(i)}$ is calculated using Orthogonal Matching Pursuit (OMP) algorithm.

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{D}^{(i)}\mathbf{X}^{(i)}\|_F^2, s.t. \|x_n\|_0 \leq T, n \in \{1, 2, \dots, N\} \tag{2}$$

Where T is the number of non-zero elements after sparse coding and N is the number of training samples. $\mathbf{D}^{(i)}$, $\mathbf{X}^{(i)}$ are the dictionary and the sparse coefficient matrix at i -th iteration.

- (c) Dictionary update: assuming that both $\mathbf{X}^{(i)}$ and $\mathbf{D}^{(i)}$ are fixed, to update the k -th column d_k of $\mathbf{D}^{(i)}$. Let the k -th row which will be multiplied by d_k in $\mathbf{X}^{(i)}$ be x_T^k , the objective penalty term can be rewritten as follows:

$$\begin{aligned} \|\mathbf{Y} - \mathbf{D}^{(i)}\mathbf{X}^{(i)}\|_F^2 &= \left\| \mathbf{Y} - \sum_{j=1}^K d_j x_T^j \right\|_F^2 \\ &= \left\| \left(\mathbf{Y} - \sum_{j \neq k} d_j x_T^j \right) - d_k x_T^k \right\|_F^2 = \|E_k - d_k x_T^k\|_F^2 \end{aligned} \tag{3}$$

In Eq. (3), $\mathbf{D}^{(i)}\mathbf{X}^{(i)}$ is decomposed into K matrices with rank 1. Assuming $K - 1$ items are fixed, the remaining k -th is the one to be updated. The matrix E_k stands for the error for all the N training examples when d_k is removed. Singular value decomposition (SVD) of E_k is conducted as follows:

$$E_k = U_k \Delta_k V_k^T \tag{4}$$

The dictionary atom d_k is replaced by the first column of $U_k (k = 1, 2, \dots, K, K < N)$.

- (d) All atoms of dictionary are updated with SVD in K times. Let $i = i + 1$, the iteration will be terminated and output \mathbf{D} if $i = I$. Otherwise go to step (b).

Each stage of sleep EEG signal can be trained into a dictionary $\mathbf{D}_i \in \mathbb{R}^{13 \times K_i}$ with above steps, then combine them into one complete dictionary $\mathbf{D} = [\mathbf{D}_{AWA}, \mathbf{D}_{REM}, \mathbf{D}_{S1}, \mathbf{D}_{S2}, \mathbf{D}_{S3}, \mathbf{D}_{S4}]$.

2.4 Classification Based on Coding Coefficients

After training dictionary, test samples can be classified by coding coefficients. For a test sample $y \in R^{13 \times 1}$ which belongs to the specific sleep stage, it could be well approximated by the dictionary D associated with the same class i using $y = Da$, which $a = [0, 0, \dots, 0, a_{i,1}, a_{i,2}, \dots, a_{i,K_i}, 0, 0, \dots, 0]^T \in R^K$ ($K = \sum K_i$) is the coding coefficient vector whose entries are zero except those associated with the i -th class and K_i is the size of i -th class dictionary. The nonzero entries in the estimate a will all associate with the columns of D from a single object class that can easily assign the test sample y to one class. The sparsest solution of $y = Da$ is defined as the following L0-optimization problem:

$$L_0 : \hat{a} = \operatorname{argmin} \|a\|_0 \text{ subject to } Da = y \quad (5)$$

The L0-minimization solution is Nonlinear Programming (NP)-hard problem. It is generally known that if just a few coefficients are not zero in vector a , the sparsest solution can be formulated as the following L1-optimization problem with an error tolerance ε :

$$L_1 : \hat{a} = \operatorname{argmin} \|a\|_1 \text{ subject to } \|Da - y\|_2 \leq \varepsilon \quad (6)$$

The sparse coding coefficient solution can be regarded as a convex optimization problem with linear matrix inequalities constraints. For classification problem, a new vector δ_i is defined, whose nonzero entries are associated with i -th class in a . Then, the test sample can be reconstructed by the coefficient of the same class. And the reconstruction residual can be calculated as follows:

$$r_i(y) = \|y - D_i \delta_i(a)\|_2$$

Finally the test sample can be classified to the specific stage with the least residual as $\operatorname{identify}(y) = \operatorname{argmin} (r_i(y))$.

We perform 2–6 stages sleep stage classification for Fpz-Oz channel and Pz-Cz channel, respectively (Table 2), testing with 10 fold cross-validations by randomly dividing all the 4318 epochs into 10 approximately equal size subsets.

Table 2. The stages included in the different number of sleep stages, 6-stage corresponds to the R&K standard.

	Sleep stages
2-stages	AWA, S1(S1 + S2 + S3 + S4 + REM)
3-stages	AWA, REM, S1(S1 + S2 + S3 + S4)
4-stages	AWA, REM, S1(S1 + S2), S3(S3 + S4)
5-stages	AWA, REM, S1, S2, S3(S3 + S4)
6-stages	AWA, REM, S1, S2, S3, S4

3 Result

3.1 The Coding Coefficient of Test Samples

For each test sample, the original 13 features can be represented by no more than 5 code coefficients by the atoms of the corresponding stage in the dictionary, while the coefficient of other stages are zero (Fig. 2).

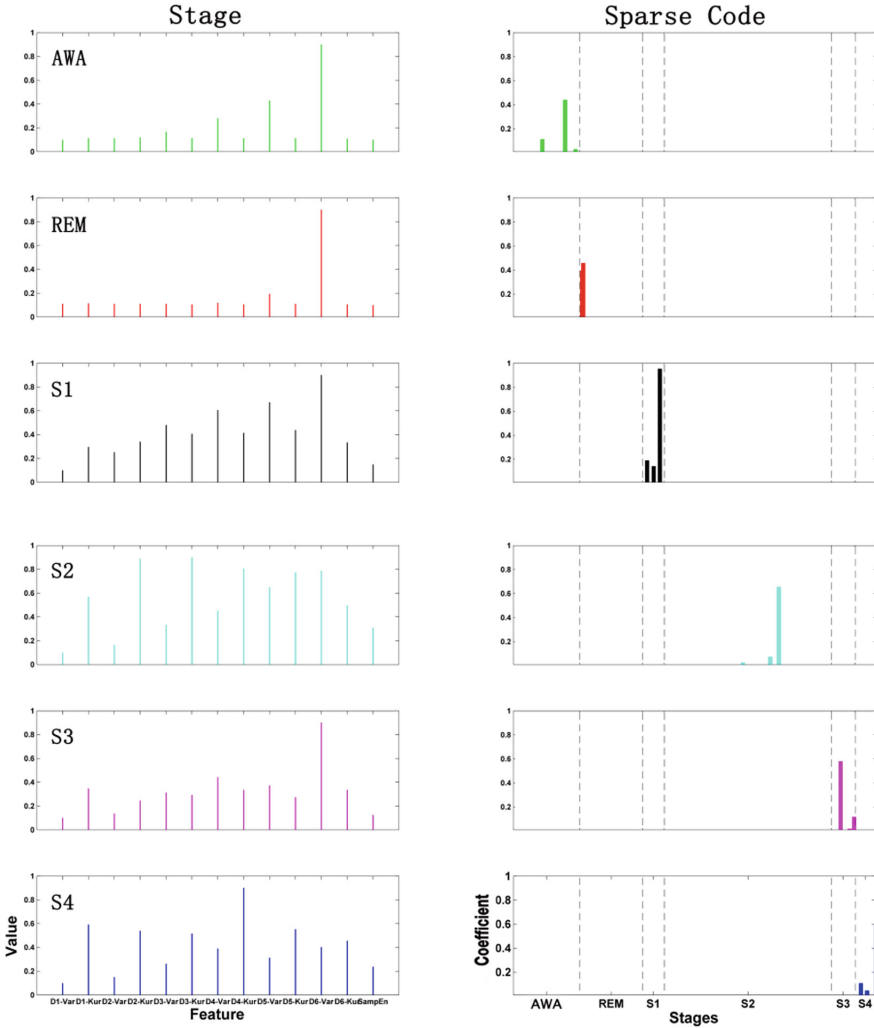


Fig. 2. The features (left) and coding coefficients (right) of each sleep stage.

For each stage, the code coefficients of test samples almost belong to the specific stage in the dictionary. For the S1 stage, some code coefficients appear in the wake and S2 stages. The overlap appeared in S3 and S4 is kind of obvious (Fig. 3).

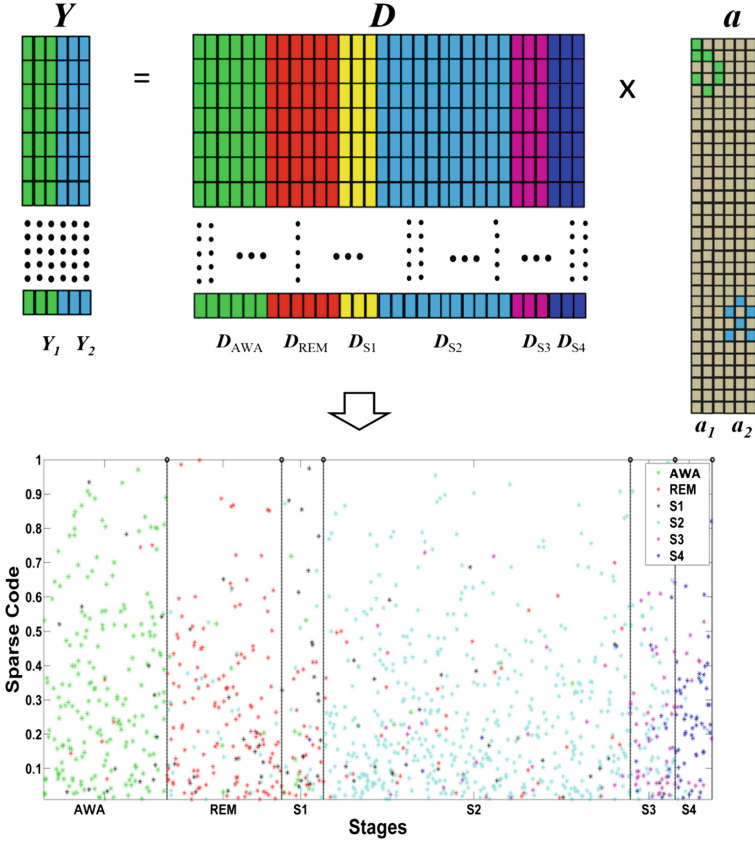


Fig. 3. The coding coefficient of whole test set. The black dotted line indicates the interface between the dictionaries.

3.2 Overall Accuracy, Precision and Recall

Table 3 shows the confusion matrix for the 6-stages sleep classification. The evaluation indexes obtained by the confusion matrix are overall accuracy, classification precision and recall rate. They are calculated as follows:

$$OA = \frac{\sum_{i=1}^Q M_{ii}}{\sum_{i=1}^Q \sum_{j=1}^Q M_{ij}}, P_i = \frac{M_{ii}}{\sum_{j=1}^Q M_{ji}}, R_i = \frac{M_{ii}}{\sum_{j=1}^Q M_{ij}} \quad (8)$$

Table 3. Confused matrix of 6-stages classification. Gray background represents Pz-Oz channel and white background represents Fpz-Cz channel

	AWA		REM		S1		S2		S3		S4	
AWA	718	621	36	22	21	28	17	106	0	11	4	8
REM	29	28	510	523	30	39	170	150	1	1	1	0
S1	35	23	73	63	67	97	91	85	2	1	1	0
S2	11	20	63	79	4	18	1845	1816	57	41	3	9
S3	2	16	2	1	0	0	87	62	142	153	56	57
S4	1	10	0	0	1	0	8	8	50	34	180	188

Where Q is the number of stages, P_i and R_i represent the precision and recall rate of i -th class, M_{ij} is the element at i -th row and j -th column in the confusion matrix.

The classification results of the two channels Pz-Oz and Fpz-Cz in different stages are compared (Tables 4 and 5).

Table 4. The classification overall accuracy (OA) comparisons of Pz-Oz and Fpz-Cz channel from 2-stage to 6-stage.

Channel	2-stages	3-stages	4-stages	5-stages	6-stages
Pz-Oz	96.52%	88.63%	85.11%	82.74%	80.17%
Fpz-Cz	93.35%	85.03%	82.96%	80.94%	78.69%

Table 5. The precision (above) and recall rate (below) of each stage for different stages classification. Gray background represents Pz-Oz channel and white background represents Fpz-Cz channel. Bold fonts indicate the better one between two channels.

	2-stages		3-stages		4-stages		5-stages		6-stages	
AWA	92.49%	89.06%	92.22%	88.07%	91.23%	87.36%	90.76%	85.60%	90.20%	86.49%
	88.08%	72.61%	89.32%	73.24%	90.09%	77.16%	91.33%	77.64%	90.20%	78.02%
REM			77.52%	74.18%	77.62%	78.34%	74.35%	76.89%	74.56%	76.02%
			66.44%	70.08%	66.94%	66.40%	69.23%	70.04%	68.83%	70.58%
S1	97.33%	94.06%	90.14%	86.88%	84.80%	82.94%	50.79%	53.85%	54.47%	53.30%
	98.38%	97.99%	94.17%	92.34%	90.41%	90.19%	23.79%	36.43%	24.91%	36.06%
S2							83.79%	82.06%	83.18%	81.54%
							91.94%	91.28%	93.04%	91.58%
S3					86.18%	82.34%	84.57%	84.08%	56.35%	63.49%
					80.15%	83.74%	83.93%	84.72%	49.13%	52.94%
S4									73.47%	71.76%
									75.00%	78.33%

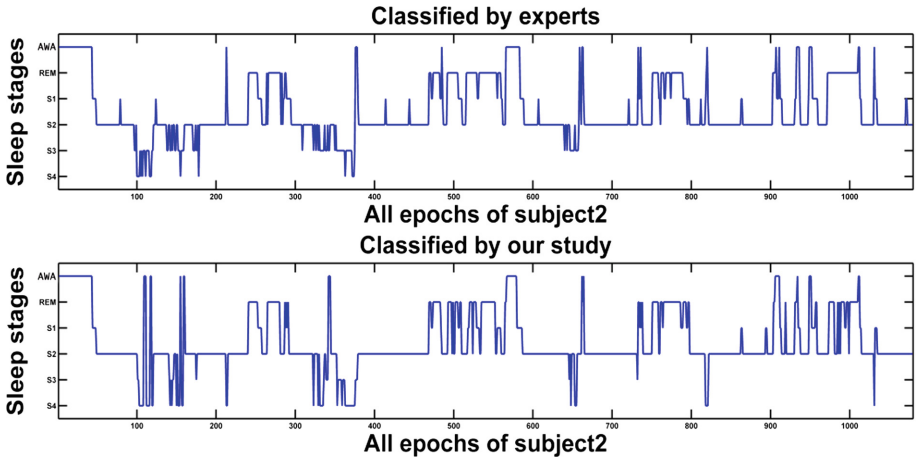


Fig. 4. Comparison of subject 2's sleep hypnogram between expert (up) and classifier (down). X-axis represents the epoch number of subject 2, Y-axis represents the different sleep stages.

Then three subjects's sleep data are used as training set, the data of another subject is test set. The classification result is given in form of sleep hypnogram (Fig. 4).

4 Discussion

Applying SRC based on K-SVD dictionary training on single-channel EEG sleep stage classification has achieved a good performance. The comparison between Pz-Oz channel and Fpz-Cz channel showed that the Pz-Oz channel was much better to be applied on single-channel EEG sleep stage classification than Fpz-Cz channel.

After sparse processing, each epoch's 13-dimensional features become no more than 5 codes. When the classification is correct, each stage can be linearly represented by the dictionary in the stage of which the test sample belongs to and the coding coefficients are non-zero in this stage while the other coefficients that not belong to the stage are zero (Fig. 2). There are a small number of misclassified samples in the wake stage and the S2 stage, so the recall rates of these two stages are relatively high (Fig. 3). The coding coefficients of all test samples showed that the misclassified samples appear mainly in REM and S2 region leading to a low accuracy because S1 stage is the transition stage between the REM stage and the S2 stage [19]. Some researchers indicate that S1 stage is too similar to the REM stage so they merged S1 stage and REM stage into one [7]. For the same reason, the S3 and S4 stages are not easy to be separated, leading that they are merged into one stage in AASM.

Compared with the state-of-art studies, the results of sparse representation for sleep stage classification showed better precision and recall rate. The 6-stage classification got the better accuracy rate compared with the literature [20] with less features, especially in S1, S2 and S4 stages (Table 5). Most of the staging results are higher than the literature. According to the sleep hypnogram of subject 2, the sleep stage classification performance obtained by SRC and experts are almost same (Fig. 4).

For Pz-Oz channel and Fpz-Cz channel, the Pz-Oz channel can get a better performance than Fpz-Cz channel for most of sleep stages in different 2–6 stages sleep stage classification especially for the wake and S2 stages (Table 5). It is not yet concluded which channel is best and still need a large amount of sample data for experimental and further study.

In summary, single channel EEG sleep stage classification can get a good performance with SRC using K-SVD dictionary training. Therefore, the use of this approach can be extended to portable sleep monitoring equipment for daily sleep monitoring. Combination of single channel EEG with other wearable electrophysiological signals (e.g. Heart rate Variation, HRV) might be helpful to improve sleep staging performance.

References

1. Chen, C., Liu, X., Ugon, A., Zhang, X., Amara, A., Garda, P., Pinna, A.: Polysomnography symbolic fusion for automatic sleep staging. In: 5èmes Journées d'Etude sur la TélÉSANTé (JETSAN) (2016)
2. Rechtschaffen, A.: A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Public health service (1968)
3. Berry, R.B., Brooks, R., Gamaldo, C.E., Harding, S.M., Marcus, C.L., Vaughn, B.V.: The AASM manual for the scoring of sleep and associated events. Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine (2012)
4. Liang, S.F., Kuo, C.E., Shaw, F.Z., Chen, Y.H., Hsu, C.H., Chen, J.Y.: Combination of expert knowledge and a genetic fuzzy inference system for automatic sleep staging. *IEEE Trans. Biomed. Eng.* **63**(10), 2108–2118 (2016)
5. Peker, M.: An efficient sleep scoring system based on EEG signal using complex-valued machine learning algorithms. *Neurocomputing* **207**, 165–177 (2016)
6. Hassan, A.R., Bhuiyan, M.I.H.: A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. *J. Neurosci. Meth.* **271**, 107–118 (2016)
7. Samiee, K., Kovács, P., Kiranyaz, S., Gabbouj, M., Saramaki, T.: Sleep stage classification using sparse rational decomposition of single channel EEG records. In: Signal Processing Conference, pp. 1860–1864. IEEE (2015)
8. Tsinalis, O., Matthews, P.M., Guo, Y.: Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Ann. Biomed. Eng.* **44**(5), 1587–1597 (2016)
9. Guo, C., Lu, F., Liu, S., Xu, W.: Sleep EEG staging based on Hilbert-Huang transform and sample entropy. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN), pp. 442–445. IEEE (2015)
10. Ren, Y., Wu, Y., Ge, Y.: A co-training algorithm for EEG classification with biomimetic pattern recognition and sparse representation. *Neurocomputing* **137**, 212–222 (2014)
11. Yu, H., Lu, H., Ouyang, T., Liu, H., Lu, B.L.: Vigilance detection based on sparse representation of EEG. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2439–2442. IEEE (2010)
12. Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Park, D.C., et al.: Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dementia* **7**(3), 280–292 (2011)

13. Shin, Y., Lee, S., Ahn, M., Jun, S.C., Lee, H.N.: Motor imagery based BCI classification via sparse representation of EEG signals. In: International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and 2011, International Conference on Bioelectromagnetism, pp. 93–97. IEEE (2011)
14. Liu, X., Shi, J., Tu, Y., Zhang, Z.: Joint collaborative representation based sleep stage classification with multi-channel EEG signals. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 590–593. IEEE (2015)
15. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Sig. Process.* **54**(11), 4311–4322 (2006)
16. Liu, F., Wang, S., Rosenberger, J., Su, J., Liu, H.: A sparse dictionary learning framework to discover discriminative source activations in EEG brain mapping. In: AAI, pp. 1431–1437 (2017)
17. Balouchestani, M., Krishnan, S.: Advanced K-means clustering algorithm for large ECG data sets based on a collaboration of compressed sensing theory and K-SVD approach. *Sig. Image Video Process.* **10**(1), 113–120 (2016)
18. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., et al.: Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation* **101**(23), E215 (2000)
19. Chen, C., Ugon, A., Zhang, X., Amara, A., Garda, P., Ganascia, J.G., Pinna, A.: Personalized sleep staging system using evolutionary algorithm and symbolic fusion. In: 2016 IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC), pp. 2266–2269. IEEE (2016)
20. da Silveira, T.L., Kozakevicius, A.J., Rodrigues, C.R.: Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain. *Med. Biol. Eng. Comput.* **55**(2), 343–352 (2017)

Artificial Intelligence and Machine Learning in Augmented Cognition



Improving Automation Transparency: Addressing Some of Machine Learning's Unique Challenges

Corey K. Fallon^(✉) and Leslie M. Blaha

Pacific Northwest National Laboratory, Richland, WA, USA
{corey.fallon, leslie.blaha}@pnnl.gov

Abstract. A variety of factors can affect one's reliance on an automated aid. Some of these factors include one's perception of the system's trustworthiness, such as perceived reliability of the system or one's ability to understand the system's underlying reasoning. A mismatch between the operator's perception and the true capabilities and characteristics of the system can lead to inappropriate reliance on the tool. This improper use of the system can manifest as either underutilization of the technology or complacency resulting from over-trusting the system. Increasing an automated tool's transparency is one approach that enables the operator to more appropriately rely on the technology. Transparent automated systems provide additional information that allows the user to see the system's intent and understand its underlying processes and capabilities. Several researchers have developed frameworks to support the design of more transparent automation. However, these frameworks may not fully consider the particular challenges to transparency design introduced by automation that leverages machine learning. Like all automation, these systems can benefit from transparency. However, artificial intelligence poses new challenges that must be considered when designing for transparency. Unique considerations must be made in terms of the type, and amount or level of transparency information conveyed to the user.

Keywords: Transparency · Machine learning · Automation

1 Appropriate Reliance

Through their interactions with automation, operators form perceptions of an automated system's technical competence, ability to function consistently (i.e., reliability) and understanding of the system's processing. These perceptions affect one's trust in the technology and ultimately one's reliance on the automation [1]. When a mismatch exists between the operator's trust and the actual capabilities of the system the operator may under-trust or over-trust the automation [1]. An operator under-trusts the automation when his or her trust is less than what is appropriate given the reliability and capabilities of the technology. For example, research on alarm systems has investigated the impact of under-trust on alarm system compliance. Several studies suggest that an unreliable alarm system which produces a high false alarm rate may result in slower and less frequent operator compliance with the system [2, 3]. Sorkin [4]

revealed that in some instances false alarms have led to a complete rejection of the technology, such as the deactivation of a warning system due to high false alarm rates.

Unreliable automation can also cause problems for the operator if the operator over-trusts the system. Over-trust can lead to complacency [1] which has been characterized by a reduction in system monitoring below what would be considered optimal, resulting in poor operator performance [5]. For example, a warfighter who over-trusts a sensor used for target detection may be less vigilant and fail to notice that the sensor is providing old or inaccurate information. In this example, the warfighter's lack of awareness would lead to inappropriate reliance on the automation (i.e., sensor) and poor performance. Complacency has been identified as one of the major factors contributing to accidents and incidents in aviation [6].

2 Automation Transparency

One possible way to calibrate trust may be to improve automation transparency [7]. In a human-system relationship transparency "is concerned with revealing information to the user and supplementing expected outputs, which reveal how a system works and/or what it is doing" [8, p. 2]. According to Lyons [9], transparency allows the operator to correctly perceive the ability, intent, and situational constraints of the automation or autonomous system. A review of the cognitive systems engineering and human factors literature identified several system design techniques to support operator cognitive performance. Techniques were identified to promote situation awareness such as providing access to historic information. Techniques for alleviating attentional demand, such as reducing visual clutter by highlighting critical information, were also identified. In addition to these techniques, the researchers included several guidelines for improving transparency such as providing operator access to unfiltered data and providing explanations for how raw data is filtered and processed. According to this review improving system transparency was identified as an important step toward supporting operator cognitive performance [10].

Empirical evidence suggests that providing operators with transparency can increase trust and operator reliance on an unreliable automated system. Fallon et al. [11] investigated the impact of Likelihood Alarm Displays (LADs) on trust. These displays generate an alarm signal coupled with additional probabilistic information regarding the signal's validity [12]. Fallon et al. [11] found that this additional transparency information significantly increased user trust in the system. In addition, Heldin et al. [13] manipulated the transparency of an automated target classification aid to support the target discrimination of fighter pilots. According to this research, providing fighter pilots with additional information about the uncertainty of the sensor increased user trust and the number of correct classifications. By improving the appropriateness of the operator's reliance on unreliable automation, the presence of transparency information may reduce errors. Also, in situations where operators have previously rejected an unreliable automated system, adding transparency to the automation may decrease operator workload by increasing reliance on the time-saving automation.

Although empirical evidence suggests increasing transparency may help calibrate operator trust and reliance on the automation, guidance for incorporating transparency

information into design is lacking. Several researchers have made some initial progress. Lyons [9] proposed multiple models for informing the implementation of transparency in human-robot interaction. For example, this researcher draws a useful distinction between what he refers to as the Intention, Task, and Analytic Models. According to Lyon's [9] Intention Model, robot designers have a responsibility to ensure that the operator fully understands the machine's functionality and purpose. In contrast, the Task Model, emphasizes the communication of system goals and progress toward those goals. The machine should also share information about its ability to perform tasks and acknowledge its errors. Finally, Lyon's [9] Analytical Model highlights the importance of sharing the system's underlying analytical processes with the operator. Adherence to this model allows the operator to understand how the system is solving problems to accomplish its goals.

Chen et al. [14] also proposed a model for designing transparency. Their model was specifically developed to support situation awareness and is known as the Situation Awareness-based Agent Transparency Model. Chen et al.'s [14] model maps onto Endsley's [15] three levels of situation awareness: perception, comprehension, and projection. Similar to Lyon's [9] Task Model, Chen et al.'s [14] model stresses the importance of providing information about system's task performance and goals. According to these researchers this type of transparency information helps facilitate operator perception. Chen et al. [14] also suggest transparency information specific to the system's underlying reasoning allows the operator to comprehend how the system is working. This type of information is consistent with Lyon's [9] Analytical Model and according to Chen et al. [14] this information promotes Endsley's second level of situation awareness: comprehension. Chen's [14] model is somewhat unique in its emphasis on projection. According to their model, transparency information should support the operators' ability to make predictions about the system's future performance.

Both Lyons [9] and Chen et al. [14] provide researchers with useful organizational frameworks from which to build. However, advances in automation in the form of artificial intelligence (AI) may require researchers to expand on the existing guidance. Specifically, automation that leverages machine learning presents new challenges and requires additional design considerations to ensure that these systems are transparent. AI that leverages machine learning can improve its reliability and accuracy with experience. The advances in automation fueled by machine learning pose new challenges for designers to ensure system transparency and appropriate reliance.

3 Machine Learning

Broadly, machine learning refers to AI systems that train and learn from past activity without the specific improvements being explicitly programmed. The systems are given algorithms for learning together with training examples/data from which they determine what to learn. There are three broad classes of learning algorithms, each with different implications for the types of interactions or user inputs that will be required or informative to the process. Because of the different level of user engagement in the learning process, each type can have different implications for the types of transparency or information that must be conveyed to those users. Supervised learning requires a

completely labeled training set from which the algorithm must draw its feedback. Traditional supervised machine learning requires all the training labels be provided up front. Advances in active and interactive machine learning are looking for ways to make this a more incremental process. In either case, the user may need a high degree of engagement with the system, implying the user must understand what the machine learner needs.

Semi-supervised and unsupervised learning algorithms need partial to no labeled input from users. This may simplify the process of constructing training exemplar sets, but it also changes the degree to which the user is engaged with the system. Unsupervised systems with minimal user interactions may also provide minimal information back to the user about the process, due to the lack of user involvement. In this case, it may be desirable to integrate corrective feedback for errors or other simple ways to engage the user, as well as training about the machine capabilities or explicit transparent information.

The increasing interest in deep learning systems has recently highlighted the impenetrable nature of black box machine learning for human observers. The hidden layers do not usually operate on human-recognizable patterns. This has resulted in a push for explainable systems, capable of translating those activities into something humans can understand. This push is based on the assumption that increasing human understanding through explanations, which are one form of transparency, will result in improved trust in the deep learning systems. What is unclear in this argument is if the explanations need to be about the internal reasoning processes or just about the classification outputs, or about some other aspect of the system entirely. Indeed, some recent work has shown that explaining the machine's reasoning can aid user in selecting the more effective classifier [16]. However, because machine learning can be used at multiple levels of automation and for multiple purposes in systems, *post hoc* explanations about machine reasoning may not always be necessary or enough to engender the appropriate trust and reliance. There are multiple types of transparency as well as degrees or levels of transparency that may be needed. Considerations of these will be informative to the system design process.

3.1 Type of Transparency

Automation equipped with machine learning adds an additional dimension of complexity and capability to the system that should be communicated to the user. At a basic level the tool should be transparent about its ability to learn and improve its own performance. This type of transparency is consistent with Lyons' [9] Intention Model. The tool should communicate to operators that it intends to learn from their input. This intention may be communicated during training or explicitly communicated to the operator during the first interaction.

In addition, the system should be clear what input is needed from the operator to improve system learning. If one of the operator's responsibilities is to teach the system, the system should provide guidance for accelerating its own learning. This is a key principle of active and interactive machine learning systems, particularly for robots, where the learner selects the data from which it will learn or requests the information or training feedback it needs [17, 18]. For example, if a recommender system based on

machine learning learns only from specific user behaviors such as user product ratings, the automation should inform the user that only this information will be helpful for learning. If the operator is not aware of this behavior's importance, he or she may choose not to provide product ratings and ultimately stunt system growth. In general, failure to communicate the importance of certain behaviors for learning can lead to inefficient or unhelpful user interactions that slow the system's progress.

Designers should develop tools that clearly communicate what user actions are required for training. However, designers should also be considerate of the training burden placed on the user and take steps to reduce this burden. It has been found that simply treating the user as an oracle and repeatedly requiring the user to give right/wrong feedback is frustrating to users [19]. This interaction puts systems into a situation where users could stop giving any input at all, thus breaking the training cycle.

A principle of mixed-initiative systems is that user interactions should be used implicitly by the system to understand user goals and provide machine support toward those efforts [20]. While mixed-initiative systems are not necessarily all machine-learning based, the same principle applies to considering how observation of ongoing user interactions and implicit learning about the user can inform the machine learning. Semantic interactions were developed as one form of implicit learning about the patterns in data of interest to the user [21]. Jasper and Blaha [22] suggested that machines might learn implicitly from interactions with user interface metaphors. We note, importantly, that while implicit learning about the user may be helpful to the system and less intrusive to the user's analytic process, they may also require some degree of transparency so that the user provides interaction inputs that are valuable to the process.

Expanding on Lyon's [9] Task Model and Chen et al.'s [14] emphasis on projection, the tool should also provide information that will help the user gauge how quickly and smoothly the system will learn as well as the upper limits of the system's performance. This type of transparency will allow a new user to manage expectations about how the system will (or will not) improve with repeated use. For some systems, the relationship between operator input and system improvement may be linear. For other automated tools system improvement may come in fits and starts despite consistent attempts by the operator to train the tool.

Appropriately calibrating user expectations with the system's rate of learning may be particularly important for tools early in their learning (i.e., novice tools). For some systems, there may be an initial training ("burn-in") period where little to no performance improvement should be expected. In other instances, the system may overcorrect early in its learning resulting in performance errors. If these learning delays are not anticipated, the user may become discouraged and underutilize or even completely reject the technology.

It may also be useful for the tool to communicate the upper limits of its capability. If it is not reasonable for the operator to expect more than 80% reliability, this information should be communicated to the operator. Without this transparency, the operators may grow frustrated when they do not see performance improve above this threshold. This is a particularly salient aspect of working with machine learning systems, because the users can often see the mistakes. If the user is expecting 100% accuracy in classification or labeling, for example, then the errors can be surprising and

unexpected or even result in a catastrophic loss of trust in the system from which the human-machine team may not recover.

For some system/environment combinations, it may be impossible for system developers to reasonably predict how quickly the system will learn or the upper limit of its performance. In these situations, it may be particularly useful for the system to provide users easy access to historical data. As Chen et al. [14] noted in their model of transparency, examining past system performance can help the operator predict future performance. AI that tracks its own performance on a task will help the operator understand how quickly the system is learning the task and the limits to the system's performance. In addition, allowing the operator to examine how much and how quickly the system learned in previous situations may provide insight into how it will learn in a new but similar environment.

3.2 Level of Transparency

The guidance above is simply an expansion of existing transparency models to support human-automation interaction [9, 14]. These models provide frameworks that organize transparency information by type such as distinguishing between task-focused and analysis-focused information. Selecting the correct type(s) of transparency information to display will help facilitate appropriate reliance and acceptance of the tool. However, designers must also consider the amount of transparency information that is appropriate for display and/or access at any given time. In this paper, we refer to the amount of transparency information as the level of transparency. We see parallels between levels of transparency and the levels of automation proposed by Parasuramen et al. [23].

The level of transparency information can range from a complete lack of transparency at the lowest level to a salient display detailing the system's performance and/or underlying reasoning process at the highest level. In addition, one might consider allowing access to detailed information within a menu structure a lower level of transparency than presenting this information on a display. Such access is consistent with Shneiderman's Visual Information Seeking Mantra broadly applicable to interactive interfaces: overview first, zoom and filter, details on demand [24, p. 365]. In practice, this means that when there is a danger of information overload, or more information than may be immediately useful to the user, the information should be made available in an easily findable way, on the demand of the user according to his or her needs. This same principle may be very useful to offering flexible degrees of transparency information, such that users desiring more details can access them, but they are not immediately cluttering the information displayed to users who do not desire the information. A variety of factors should be considered when choosing the appropriate level of information. Some of these factors include the operator's workload and experience with the tool, the reliability of the automation, and consequences of an error. Choosing the appropriate level of transparency is an important design decision for any automated system. However, in this paper we will focus specifically on the unique challenges posed by automated tools that improve with user interaction.

In static automated systems, the required level of transparency for appropriate reliance may drop as the user learns the capabilities of the tool and how it processes information. In the case of machine learning, increased familiarity with the tool coupled

with improved performance may require the need for very little transparency. However, it is important to consider additional factors when deciding the level of transparency. One important factor to consider is the consequence of an error in the operational environment.

In an environment with relatively low stakes, a user who is familiar with the highly reliable automated tool may have little need for transparency. A high level of transparency information under these conditions may at best be ignored and at worst be a distraction that places unwanted attentional demands on the operator. Perhaps the ideal design approach for a low consequence environment is an adaptive display that reduces its level of transparency as both system reliability and operator familiarity increase. At peak human-machine teaming performance, the operator interacts with the automation seamlessly with little explicit communication. This relationship is similar to a high functioning human-human team that relies on implicit coordination [25, see Fig. 1].

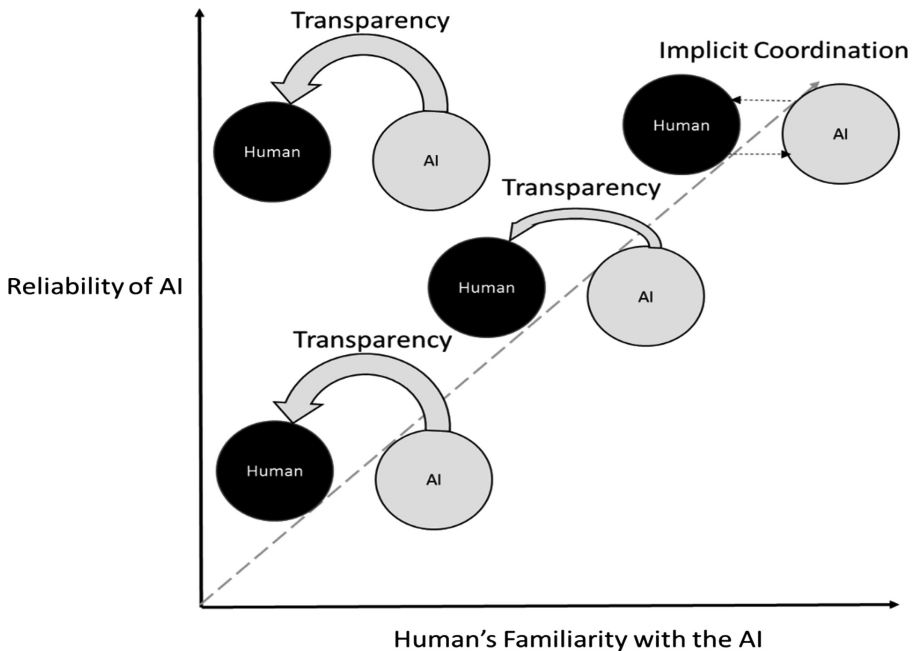


Fig. 1. The impact of operator familiarity and AI reliability on the level of AI transparency.

Figure 1 also depicts a scenario where the automation's reliability is high despite the user's lack of familiarity with the tool. Such a situation might exist if the AI has been trained by other users and is now available for a new operator to use. We propose that the level of transparency should be high in this situation despite the system's high reliability. A high level of transparency may be necessary because the user lacks the hands-on experience needed to become aware of the system's superior performance. The high level of transparency can be used to accelerate trust calibration and appropriate reliance in the absence of familiarity.

It is important to note in certain situations it may never be appropriate to design for very low levels of transparency. In high stakes environments where the consequences of an error are severe, it may always be appropriate to provide a high level of transparency. For example, despite familiarity with a highly reliable decision support tool, a high consequence environment may still require the need to understand the reasoning behind the system’s recommendation [see Fig. 2].

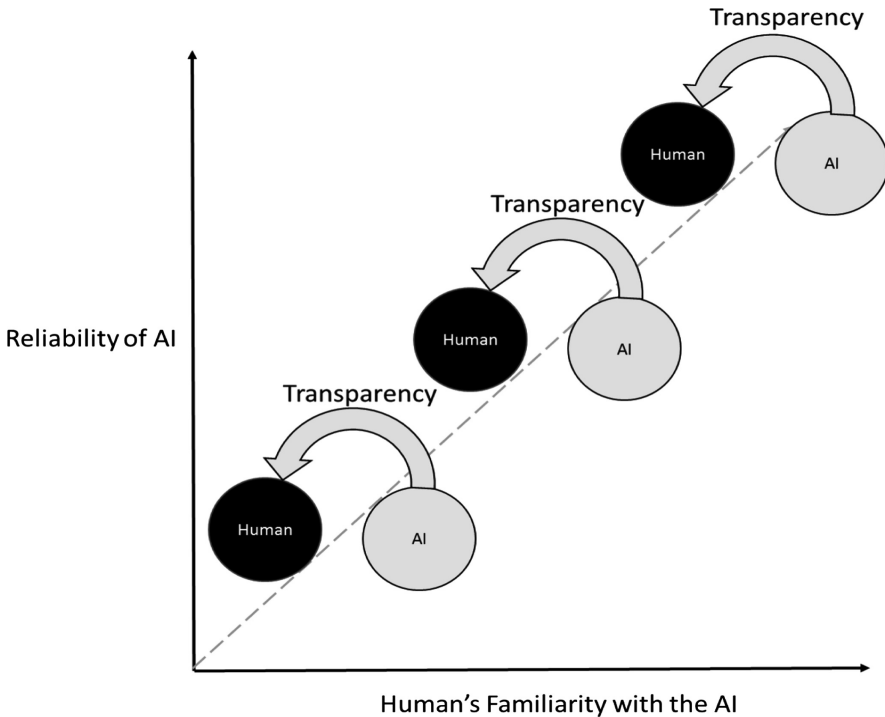


Fig. 2. The impact of operator familiarity and AI reliability on the level of AI transparency in a high consequence environment.

4 Conclusion

The success of a human-machine team is dependent on an operator’s ability to appropriately rely on the automation. Over-reliance can lead to complacency, lapses in operator attention and error. Underutilization of the tool can result in increased workload and inefficiencies. Through trial and error the operator may be able to calibrate his or her trust in the tool and learn to rely on it appropriately. However, the trial and error technique can be time consuming and prone to errors as the operator attempts to understand the system’s capabilities and limits. In addition, trial and error may never fully reveal the underlying analysis that governs the system’s behavior. Increasing system transparency may be one technique to facilitate appropriate reliance more efficiently and with fewer errors.

As the capabilities of automated tools grow, the added complexity of these tools poses new challenges for transparency design. Machine learning in particular adds an additional dimension that must be considered when building transparency into these systems. This advancement in AI requires designers to consider a new type of transparency. In addition to communicating information about the systems' capabilities, goals and underlying reasoning, automation that leverages machine learning should communicate information about how the system learns.

Machine learning also creates additional challenges for designing the appropriate level of transparency. Systems governed by machine learning will likely improve with operator interaction. In low consequence environments, this increase in system accuracy and reliability may benefit from a decrease in transparency level. A level that adjusts automatically or is adjustable by the operator may be ideal given the potential for shifts in system performance over time.

Acknowledgments. This effort was sponsored by the Analysis in Motion Initiative at the Pacific Northwest National Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

1. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. *Hum. Factors* **46**, 50–80 (2004)
2. Bliss, J.P., Fallon, C.K.: Active warnings II: false warnings. In: Wogalter, M.S. (ed.) *The Handbook of Warnings*, pp. 231–242. Lawrence Erlbaum Associates, Mahwah (2006)
3. Getty, D.J., Swets, J.A., Pickett, R.M., Gonthier, D.: System operator response to warnings of danger: a laboratory investigation of the effects of the predictive value of a warning on human response time. *J. Exp. Psychol. Appl.* **1**, 19–33 (1995)
4. Sorkin, R.D.: Why are people turning off our alarms? *J. Acoust. Soc. Am.* **84**(3), 1107–1108 (1988)
5. Parasuraman, R., Manzey, D.H.: Complacency and bias in human use of automation: an attentional integration. *Hum. Factors* **52**, 381–410 (2010)
6. Funk, K., Lyall, B., Wilson, J., Vint, R., Niemczyk, M., Suroteguh, C., Owen, G.: Flight deck automation issues. *Int. J. Aviat. Psychol.* **9**(2), 109–123 (1999)
7. Fallon, C.K., Murphy, A.K.G., Zimmerman, L., Mueller, S.T.: The calibration of trust in an automated system: a sensemaking process. Paper published in *The 2010 International Symposium on Collaborative Technologies and Systems*, Chicago, IL, pp. 390–395 (2010)
8. Osofsky, S., Sanders, T., Jentsch, F., Hancock, P., Chen, J.Y.C.: Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In: *SPIE Defense + Security*, pp. 90840E–90840E-12 (2014)
9. Lyons, J.B.: Being transparent about transparency: a model for human-robot interaction. In: Sofge, D., Kruijff, G.J., Lawless, W.F. (eds.) *Trust and Autonomous Systems: Papers from the AAAI Spring Symposium (Technical report SS-13-07)*. AAAI, Menlo Park (2013)
10. Long, W., Cox, D.A.: Indicators for identifying systems that hinder cognitive performance. In: *Proceedings of the Eighth International Conference on Naturalistic Decision Making*, Asilomar, CA, pp. 171–175 (2007)

11. Fallon, C.K., Bustamante, E.A., Ely, K.M., Bliss, J.P.: Improving user trust with a likelihood alarm display. In: Proceedings of the 11th International Conference on Human-Computer Interaction, Las Vegas, NV (2005)
12. Sorkin, R.D., Kantowitz, B.H., Kantowitz, S.C.: Likelihood alarm displays. *Hum. Factors* **30**, 445–459 (1988)
13. Helldin, T., Ohlander, U., Falkman, G., Riveiro, M.: Transparency of automated combat classification. In: Engineering Psychology and Cognitive Ergonomics, pp. 22–33 (2014)
14. Chen, J.Y.C., Procci, K., Boyce, M., Wright, J., Garcia, A., Barnes, M.: Situation Awareness-Based Agent Transparency. (Final Report, Army Research Laboratory 6905) Aberdeen Proving Ground, MD 21005-5425 (2014)
15. Endsley, M.R.: Toward a theory of situation awareness. *Hum. Factors* **37**(1), 32–64 (1995)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)
17. Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the people: the role of humans in interactive machine learning. *AI Mag.* **35**(4), 105–120 (2014)
18. Guillory, A., Bilmes, J.A.: Simultaneous learning and covering with adversarial noise. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 369–376. International Machine Learning Society, Inc., Princeton (2011)
19. Cakmak, M., Chao, C., Thomaz, A.L.: Designing interactions for robot active learners. *Auton. Ment. Dev.* **2**(2), 108–118 (2010). <https://doi.org/10.1109/TAMD.2010.2051030>
20. Horvitz, E.: Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 159–166. ACM (1999)
21. Endert, A., Fiaux, P., North, C.: Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Trans. Vis. Comput. Graphics* **18**(12), 2879–2888 (2012)
22. Jasper, R.J., Blaha, L.M.: Interface metaphors for interactive machine learning. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) AC 2017. LNCS (LNAI), vol. 10284, pp. 521–534. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58628-1_39
23. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **30**(3), 286–297 (2000)
24. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: The Craft of Information Visualization, pp. 364–371. Morgan Kaufmann, Amsterdam (2003). <https://doi.org/10.1016/B978-155860915-0/50046-9>
25. Espinosa, A., Lerch, J., Kraut, R.: Explicit vs. implicit coordination mechanisms and task dependencies: one size does not fit all. In: Salas, E., Fiore, S.M., Cannon-Bowers, J.A. (eds.) Team Cognition: Process and Performance at the Inter- and Intra-individual Level (2002). <https://doi.org/10.1037/10690-006>



Artificial Intelligence for Advanced Human-Machine Symbiosis

Scott S. Grigsby^(✉)

Soar Technologies, Inc., Ann Arbor, MI, USA
scott.grigsby@soartech.com

Abstract. Human capabilities such as memory, attention, sensory bandwidth, comprehension, and visualization are critically important but all have innate limitations. However, these human abilities can benefit from rapidly growing computational capabilities. We can apply computational power to support and augment cognitive skills that will bolster the limited human cognitive resource and provide new capabilities through this symbiosis. We now have the ability to design human-computer interaction capabilities where the computer anticipates, predicts, and augments the performance of the user and where the human supports, aids, and enhances the learning and performance of the computer. Augmented cognition seeks to advance this human-machine symbiosis through both machine understanding of the human (such as physical state sensing, cognitive state sensing, psychophysiology, emotion detection, and intent projection) and human understanding of the machine (such as explainable AI, shared situation awareness, trust enhancement, and advanced UX). The ultimate result being a truly interactive symbiosis where humans and computers are tightly coupled in productive partnerships that merge the best of the human with the best of the machine. As advances in artificial intelligence (AI) accelerate across a myriad of applications, we seek to understand the current state-of-the-art of AI and how it may be best applied for advancing human-machine symbiosis.

Keywords: Artificial intelligence · Human-machine teaming
Augmented Cognition · Human-machine symbiosis · Situation awareness

1 Introduction

Humans have been seeking ways to perform work easier and better since the first instance stone hit flint. All tool use is at its basis an instance of a human and machine (even a simple machine like a wedge or screw) interacting to make a job easier. A hammer can apply more force to a given area than a human could alone but the hammer is useless without the human to wield it. In modern times, our tools have become more and more complex to work with humans to perform ever more difficult jobs (or perform simple jobs more effectively). A modern washing machine works with a human to make the job of washing clothes much easier. The human uses their skills to transport, sort, and load the machine, then the machine uses its skills to sense dirt level, load level, adjust water height and temperature, and repeat repetitive actions

(something machines are especially good at) to agitate the clothes and then spin the excess water out. This pairing of man and machine makes the entire process easier and faster - a prime example of the benefits of humans and machines working together to improve the system's ability to perform *physical* work. In the cognitive domain, the use of computers and basic software is a further example of machines and humans working together to aid *mental* work. The power of a simple spreadsheet comes from the pairing of a human - who can set up and initialize data sets, with a machine - that can perform rapid accurate calculations and store them in memory indefinitely - to allow the human to easily transform, visualize and share data. Through millions of years of evolution, humans have evolved an innate capacity for intuition, analogy, creativity, and induction to ask complex questions, but we still struggle and require years of schooling to master computation and logic, and require rote repetition to develop large semantic memory stores (what's the capitol of Wyoming?), areas where machines excel to quickly and accurately help answer our questions.

While something like a spreadsheet is a simple example of humans and machines working together, what many imagine now is a next step, where humans and advanced intelligent machines team together seamlessly and symbiotically to perform ever more complex cognitive and physical tasks. To do this, requires the development of machines that think and understand, at least to a level where they can understand and anticipate human actions and intent, and communicate this understanding.

As with human cognition and human teams, a primary aspect for success is context. Successful decisions and actions are not made in a vacuum but require knowledge of the current environment - what has happened in the past as well as future possible actions and outcomes. Ask any experienced team leader to speculate on a decision they may make in the future and they will usually respond "It depends." It depends on their situation awareness of the world they are acting in, on the abilities and state of the people and systems they are interacting with, and on their ability to trust and understand what those people and systems may do. Within the human-computer symbiosis paradigm, these aspects are equally important for success. To aid researchers, designers, developers, and potential users of these systems, we need to understand how humans and machines may team together in the future - how machines can be made to start to understand and adapt to context through advanced artificial intelligence techniques, how they can use this to develop a shared situation awareness with their human teammate, how they can start to learn about their human teammate's emotional and cognitive state, and the importance of developing trust between man and machine in order for these systems to be successful.

2 Artificial Intelligence and Human-Machine Teaming

Since the first imagining of so-called thinking machines or artificial intelligence (AI), there has been much debate as to whether machines would, could, or should supplant humans or enhance humans. In theory, intelligent machines could be designed to do either. These systems have been differentiated as "cognitive prostheses" - designed to replace human capability vs "cognitive orthotics" - designed to enhance and add to human capabilities [1]. While even a prosthetic system requires some level of

interaction, orthotic systems are designed from the beginning to be a tool for direct human teaming. As Nirenburg states:

“Orthotic systems... are intended to collaborate with humans on carrying out tasks, serving as high-functioning members of a society populated by a mixture of humans and artificial intelligent agents. As the intelligent agents in this society, orthotic systems must both perform tasks and communicate at a human level.”

These orthotic systems are synonymous with our human-machine symbiosis paradigm where humans and machines work seamlessly within the same world model to understand and solve problems.

A qualitative example of the difference between prosthetic and orthotic implementations of a technology is in the use of facial recognition. Facial recognition technology within Facebook acts independently of a human operator to detect and identify faces within images posted online. Even though the output may eventually be used by a human to, say, annotate pictures of relatives, once built, the systems itself no longer requires human interaction for it to perform its task. In fact, except by the programmer, there is no way for a human to interact and change how the system processes faces. It is simply a tool or prosthetic for the human. However, if we incorporate the same facial recognition technology into an augmented reality display, we can track the faces within a crowd and automatically bring up information about an acquaintance we are talking to that could help us interact with them. Here, the system acts as an orthotic display. If we also paired this real-time facial recognition technology with emotion detection algorithms that could detect the friend’s emotional state, and combined that with measures of our own emotional state (through physiological or voice stress sensing, and language understanding), the system may actually be able to detect and help diffuse an ensuing argument. In addition, in real time or in review, we could provide feedback to the machine’s situational assessment system allowing it to learn subtle nuances of human-human interaction to improve its algorithms – a symbiotic interaction between the machine and user making each better.

The core to these systems lies in the underlying intelligence of the software and design of architectures to support shared situation awareness and a common world model and goals – the world of artificial intelligence.

2.1 1st Wave AI – Handcrafted Systems

To date, artificial intelligence (AI) has developed many useful prosthetic tools and techniques for performing very specific functions. However, as complex as these programs and expert systems can be, they are very brittle and must follow specific rules and logic and be written line by line by human designers. This is the so-called handcrafted AI, now called “1st wave AI” [2]. Even a standard simple computer program can be considered to fall into this category as they are designed and coded to perform a specific purpose. For example, a C program that does a simple bubble sort (Fig. 1) could still be considered a rudimentary form of AI as it is replacing what a human brain would do with a computer algorithm - a first step in emulating human cognition.

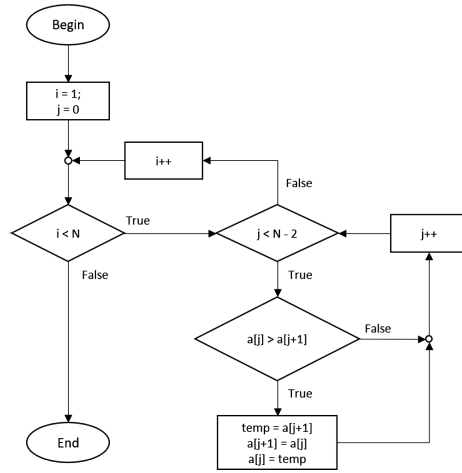


Fig. 1. Flow chart for a simple bubble sort

The key aspect of 1st wave AI is that programmers first figure out how to solve the problem, then they code the solution in a step-by-step manner. This can get very complex, but it is still the encoding of a solved problem. Our feeling that these programs do not count as AI may reflect the so-called “AI Effect” [3] that any problem that has been solved is no longer considered difficult and therefore is no longer in the realm of AI. As Brooks states “Every time we figure out a piece of it, it stops being magical; we say, ‘Oh, that’s just a computation.’” [4] But a system that can automatically take a large random data set and organize it is certainly a time saving tool that fits within the paradigm of being a tool for helping and replacing human mental work and is at its base level a form of AI. However, the inherent limits of 1st wave AI and the desire for systems that show more emergent intelligence have led to the next step of AI.

2.2 2nd Wave AI – Statistical Systems

The 2nd wave of AI is currently in its heyday and consists of statistical systems that use powerful new techniques and computing resources to perform statistical object recognition and look for patterns in large batches of data. Originating with neural network algorithms first developed in the 1950s [5], these systems originally sought to mirror human brain organization by mimicking the neural connectivity of neurons and synapses at a very simplistic level. Within neural nets (also called connectionist models), “neurons” have a simple input/output structure and perform one layer of processing between each subsequent “synaptic” connection layer. The analogy with human brain systems is so tenuous however, that most system practitioners now talk more of nodes and layers as opposed to neurons and synapses.

With the development of the latest deep neural net (DNN) algorithms and increased computing power, these systems have become powerful enough to recognize categorical objects (“cat”, “car”) and faces (Fig. 2) in images, beat human chess masters, and surpass humans at language recognition (however not language understanding).

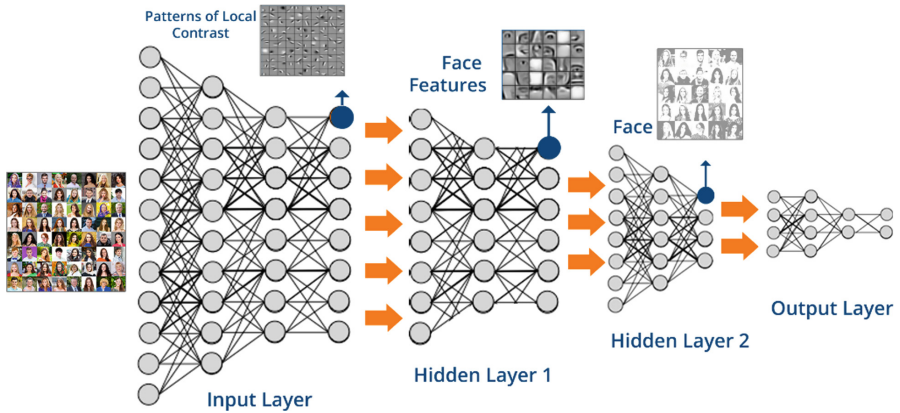


Fig. 2. Structure of a deep neural net (DNN) for facial recognition [6].

While extremely powerful, there are several drawbacks to these statistical systems. First is the requirement of tens of thousands of examples for the system to learn. Second is the limited ability of the system once trained to only perform specific tasks. For example, a system trained to recognize a cat in images can do quite well at recognizing cats but will fail completely if suddenly asked to recognize a maple tree. A chess playing bot that could beat a Grand Master would be unable to perform the first move in Go. These systems have no general or real-time interactive learning ability, they cannot work with dynamic goals and context, nor cope with abstract reasoning or language comprehension. In addition, because of their statistical nature, these systems do not easily lend themselves to explanation. The systems develop a black box solution to their specific problem making it extremely difficult to understand and debug errors, for example. While there are large efforts to try to develop workarounds and solutions to this explain-ability issue, an alternative approach lies in the so called 3rd wave of AI which is designed to supplant the issue altogether and develop machine-based cognitive reasoning systems more analogous to human reasoning and capabilities.

2.3 3rd Wave AI – Cognitive Architectures

The 3rd wave of artificial intelligence designs systems that use contextual adaptation – systems that, like humans, can reason, learn, and adapt to their environments. Third wave AI is founded on the *cognitive systems paradigm* [7] which distances itself from the current mainstream statistical AI and seeks to understand the fundamental processes of how cognitive systems - humans and machines - can reason, learn, and explain.

A main focus on the development of cognitive systems is on the development of cognitive architectures that form a framework for the basic components and connectivity of the various modules. An architecture can be thought of as a fixed set of structures and mechanisms or processes. Complex systems can be decomposed into an architecture and its content which, when combined, results in behaviors. Cognitive behaviors have certain aspects: they are goal-oriented, reflect a rich complex

environment, require a large amount of knowledge, require the use of symbols and abstractions, are flexible, and require experience and learning [8].

Cognitive architectures are theories of fixed mechanisms and structures that underlie this cognitive behavior – human or otherwise. These cognitive architectures form the blueprint for the development of software intelligent agents to solve higher level problems. This blueprint consists of “its representational assumptions, the characteristics of its memories, and the processes that operate on those memories” [9].

Laird et al. [10] have recently published a Standard Model of the Mind that seeks to formally unify the components and processing design of all human-like cognitive systems be they based in AI, neuroscience, or robotics (Fig. 3).

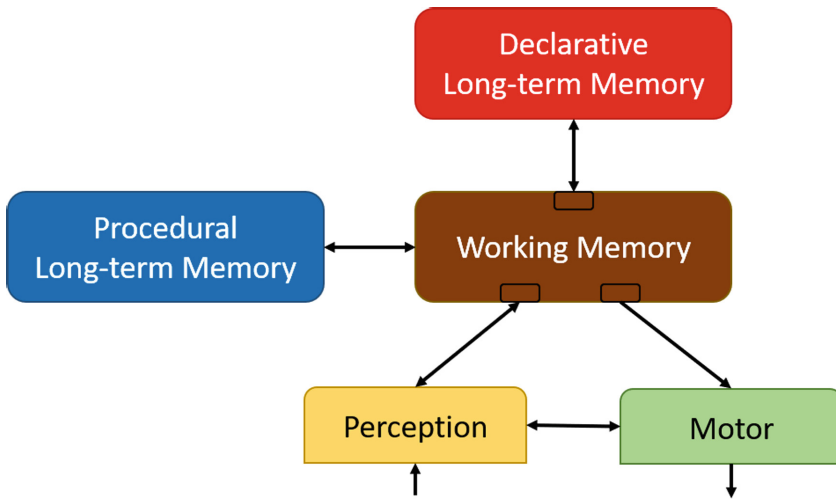


Fig. 3. The standard model of the mind [10].

This standard model integrates components of memory, perception, and action with processes for learning, communication, and representation. The core fundamentals of the Standard Model are:

- (1) Processing yields bounded rationality not optimality
- (2) There is significant parallel processing (within and across modules)
- (3) Complex behavior arises from a sequence of independent cognitive cycles that operate in their local context
- (4) Declarative and procedural long-term memories contain symbol structures and associated quantitative metadata
- (5) Global communication is provided by short-term memory: rule-like conditions and actions exert control by altering the contents of working memory
- (6) Global control is provided by procedural long-term memory
- (7) Factual knowledge is provided by declarative long-term memory
- (8) Long-term memory content is learnable

- (9) Procedural learning involves at least reinforcement learning and procedural composition
- (10) Declarative learning involves the acquisition of facts and tuning of metadata
- (11) Perception yields symbol structures with associated metadata in specific working memory buffers
- (12) Attentional bottleneck constrains information available in working memory
- (13) Perception can be influenced by top-down information from working memory
- (14) Motor control converts symbol structures to external actions
- (15) There can be multiple motor modules

Within this standard framework, however, there is some diversity in the approaches and goals of different cognitive architectures and models. Some approaches, like ACT-R [11], seek to directly model and re-create how the human mind works in order to study and learn from it. However, other cognitive architectures, like Soar [8], seek to re-create and optimize these human abilities and capabilities within the computer without necessarily directly modeling exactly how the human may do it (analog as opposed to homologs). A deeper dive into the design of the Soar architecture shows the advantages of these systems for developing the requisite contextual adaptation needed for human-machine symbiosis.

Soar is designed to have a basis in cognitive behavior that lends itself to a shared awareness with humans. It is optimized to support knowledge-intensive reasoning, hierarchical reasoning, planning, and reactive execution. Soar is grounded in Newell's *unified theory of cognition* [12] and incorporates decades of experience in cognitive research into an architecture that is reusable across new cognitive models. The Soar architecture assumes cognitive behavior is goal-oriented and that behavior is a reflection of movement through a problem space to make decisions to reach the required goal. It is both a theory of what cognition is and a computational implementation of that theory. Soar has features that mimic human cognition such as perception, working memory, procedural memory, semantic memory, episodic memory, reinforcement learning, and decision making. Soar deals well in complex environments by having an ability to apply algorithms based on the context of the situation and the type and quality of the data. Additionally, Soar's primary advantage is that it is designed for *satisficing* - trying to find an acceptable option based on what it knows. The architecture is built to work even when dealing with incomplete data, when data is out of order, or is unexpected. This ability gives Soar (and cognitive architectures in general) a major advantage over standard math-based or rigid AI approaches for human-machine teaming applications.

However, regardless of the specific cognitive system architecture that may eventually be used for human-machine symbiosis, a main challenge for any interaction (whether human-human or human-machine) is the coordination of intent, expectations, and effects across actors so as to achieve one or more common goals. Characteristics like trust, shared situation awareness, and intent recognition will be key to this endeavor.

3 Situation Awareness

In 2012, the Defense Science Board (DSB) recognized a gap in autonomy research, determining that research was focusing far more on the computer and levels of autonomy than on useful design principles that support the human-system interaction and dynamic function allocation between human and machine [13]. To achieve any level of human-machine symbiosis, the human and the machine must share understanding and situation awareness (SA) and adapt to the needs and capabilities of the other.

Teaming of any sort requires the need to maintain SA of many relevant aspects of task performance and outcomes and to achieve an allocation of roles, functions, behavioral capabilities, and resources which serves the integrated team. This includes:

- (1) Staying aware of all important activities and events
- (2) Predicting outcomes and activities so as to proactively address them
- (3) Making decisions about how to best act or react
- (4) Shifting and regaining awareness when multiple activities occur simultaneously
- (5) Triaging and offloading tasks when overloaded
- (6) Recognizing faults or failures (within the system and within the process)

For human-machine teaming applications, we adopt the standard definition of situation awareness as “the **perception** of the elements in the environment within a volume of time and space, the **comprehension** of their meaning, and the **projection** of their status in the near future.” [14] However, human-machine teaming adds the complexity of requiring both the human’s SA of the world (including knowledge of the state of the autonomous teammate) and also the machine’s SA of the world (including knowledge of the state of the human operator) with the ideal being a state of shared situation awareness between human and machine.

For human-machine symbiosis, the advantage of cognitive systems and cognitive architectures is that they have built in analogs to the standard SA processes of perception (or sensing), comprehension (or reasoning), and projection (or planning). Using these components, the cognitive system can develop its own model for situation awareness in the form of an internal operating picture (IOP) which monitors goals and objectives, tasks, operator and system states, and sensor feeds with the goal of developing a computational “mental” model of an expert’s SA within the autonomous system [15].

While only a first step towards the level of shared SA required for human-machine symbiosis, the IOP that emerges within cognitive systems is based on the in-memory store of separate indexed data structures for things like: objects in the world, tasks, goals and objectives, and other relevant state items. Once developed, the IOP is maintained by a set of processes that are continuously monitoring incoming state messages from the environment and either simply taking them in and storing them (perception), making additional computational inferences on them (comprehension), or planning for future action and events (projection).

The ability for an intelligent machine to maintain an internal operating picture can form the basis for situation awareness for both human users and the machine. Within a

decision support application, the system could output its “knowledge” using a smart interaction module (such as a display or natural language communication) to provide alerts or events, allocate tasks, or otherwise inform the user, allowing the human user to take intelligent actions. The IOP also provides a mechanism for the machine to take in information about the human user to start to understand the state of the human operator and allow the system to adapt as necessary to the human’s cognitive and emotional state.

4 Human State Sensing

The growth of “affective computing” (devices to discern human emotions) holds promise to revolutionize computing interaction by engaging users at a deeper level. During human-human interaction, much information is passed between people through body language, tone, affect, facial expression, and more, to relay valuable information about the human’s state emotional state and ability to cope with a situation. Standard human-machine interaction (like keyboard and mouse or even current speech systems), lack this ability to discern the user’s emotional state and glean valuable information about how to interact with the user.

In the early 2000s, DARPA’s Augmented Cognition (AugCog) program saw a major push to incorporate human cognitive state sensing within computer interfaces to provide “order of magnitude increases in available, net thinking power resulting from linked human-machine dyads.” [16] The ultimate goal of AugCog, and operational neuroscience programs like it, is a complete human-machine symbiosis where the human understands and intuitively reacts to the machine and the machine understands and “intuitively” reacts to the human. In order for this level of interaction to be achieved, systems will need to be developed to recognize human stress responses (e.g., cognitive load) and emotions.

The core of current affective computing systems is the use of speech intonation and facial expressions to discern human emotion. While functional for general use (such as improving tutoring systems), these measures are not robust enough for many applications which need systems that are connected to the operator and sense emotional and physiological responses discreetly and, ideally, before the operator has an outward response. This type of interaction would allow teaming with machines that can sense and react in real-time before emotions cause issues to be avoided or rectified.

Physiological sensing may be able to provide this capability but to date it’s use for real time sensing has been problematic. The accurate determination and measurement of psychophysiological constructs requires a thorough understanding of physiological signals and their cognitive correlates. The fundamental challenge for quantitative, sensor-based cognitive and emotional modeling is the complexity of the underlying physiology. Several distinct physiological processes influence the physiological responses and the various external sensing systems may pick up different elements or admixtures of stress, workload, and/or emotional response in varying degrees making it hard to isolate any one construct.

Workload itself has been studied for decades and has shown it can be a valuable measure of human state for computer interaction [17, 18]. The primary goal of

workload assessment within human-computer interaction to date has been to use these signals to inform when to mitigate workload effects by either offloading tasks or changing the interaction in some way. However, some users can perform under a high workload with little stress or reaction (or even a positive sense of challenge) while others may start to have adverse emotional reactions even though their workload remains fairly low. For nuanced interaction within a human-machine symbiosis paradigm, the machine should be able to sense not just high workload, but the human's emotional state. Emotion is a primary aspect for conveying and understanding human reactions to events and there is a strong relationship between human emotional episodes and the way humans subsequently think and act.

The key to reliable measures of workload, strain, and other human state sensing may lie in the development of an intelligent AI-based signal classifier and cognitive system-based reasoner based on 3rd wave AI architectures. These system would provide the ability to take in a variety of psychophysiological signals, understand the context and environment that is leading to the generation of those signals, and reason across incomplete complex data sources to understand the user state. This capability would not only allow for better individualized and context-based signal classification, it would form a direct input to the machine's IOP and correlate the human state response to the task criticality and the user's ability to adequately perform the task given an emotional response. This type of classifier would be a valuable construct for aiding machine awareness of user state and allowing it to adapt and respond more intuitively to user actions and needs.

5 Trust in Autonomous Systems

Even with properly designed 3rd wave architectures, IOPs for situation awareness, and machine sensing of human state, decision making between teams of humans and automated systems or agents invokes all of the problems and opportunities of human-automation interaction in general, but adds dimensions of close coordination and behavioral unpredictability. Humans use shared training, culture, natural language (including jargon, affect, "body language", etc.) and the simple fact of sharing human-centric expectations ("common sense") to make this coordination tractable and yet flexible. Even then, humans have developed particular forms, protocols, heuristics, and procedures to enhance coordinated decision making and trust. The challenge for human-machine symbiosis is to achieve at least this degree of trust and shared expectations with minimal communication overhead, while still preserving the computational speed and precision that machines afford. Within the human-machine teaming paradigm, trust can be defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" [19].

The goal is not necessarily to achieve 100% operator trust in an autonomous system, but rather to achieve the right level of trust in the system: trust that is appropriately calibrated to accurately reflect system intent, performance level, and context. Excessive trust in a non-perfect system will lead to over-reliance and error.

Excessive distrust in a system will result in disuse and increased human tasking. Appropriate or “calibrated” trust, will only be achieved reliably when transparency is established between man and machine where transparency is defined as the extent to which the intent, ability, and constraints of the autonomy are accurately perceived. The degree to which we can achieve calibrated human-machine trust is dependent upon aspects of the human, the context, and the system.

The challenge is to identify the antecedents of trust and to develop usable, empirically-derived guidelines for future systems that will help foster human-machine trust calibration through improved autonomous system designs and human-machine team training approaches. These factors are widely varied, with some being attributes of the user (e.g., age, gender, culture, experience, biases), some being attributes of the system (e.g., accuracy, reliability, error types, communication modes/styles), and some being attributes of the human-system task context (e.g., task complexity, risk, workload). Further, these factors are embedded within a construct of human-machine transparency, which is essential to achieving calibrated trust. With his focus on achieving calibrated human-machine trust, Lyons [20] has introduced a model of human-machine transparency that reflects several dimensions of the human-machine teaming relationship, including system intent, system goals and tasks, analytic principles, environment or context, division of labor, and system understanding of the human state. This paradigm may provide the scaffolding for the trust relationship required for human-machine symbiosis in the future.

6 Conclusions

The goal of human-machine symbiosis is to design physical and cognitive devices that enhance and add to human capabilities. For this goal, systems need to be “intelligent” and capable of understanding humans but do not necessarily need to re-create how humans think. While direct human modeling homologs may enhance the machines ability to understand, it is not a necessary condition of human-machine symbiosis where analogs may prove equally valuable. Intelligent machines based on neural networks and other machine learning techniques are powerful tools but just like screwdrivers and automobiles, they are designed to help humans solve very specific goals. Systems for human-machine symbiosis will need much broader capabilities in that they seek to aid humans across a wide swath and variety of dynamic problem solving in a seamless way. For this advanced interaction, a more general artificial intelligence, such as those grounded in cognitive systems, will have to be a core attribute of the machine. That is not to say that these systems should be built only around cognitive architectures. It is easy to imagine that the best systems may be hybrids, utilizing neural networks for aspects of vision and classification, combining them with strict rules-based expert systems for doing rote actions, and placing them under the auspices of higher level cognitive reasoning systems. The ultimate result being a truly interactive symbiosis where humans and computers are tightly coupled in productive partnerships that merge the best of the human with the best of the machine.

References

1. Nirenburg, S.: Cognitive systems: toward human-level functionality. *AI Mag.* **38**(7), 5–12 (2017)
2. Launchbury, J.: A DARPA Perspective on Artificial Intelligence (2017). https://www.youtube.com/watch?time_continue=5&v=-O0IG3tSYpU
3. McCurduck, P.: *Machines Who Think*, 2nd edn. AK Peters Ltd., Natick (2004). ISBN 1-56881-205-1
4. <https://www.wired.com/2002/03/everywhere/>
5. Kleene, S.C.: Representation of events in nerve nets and finite automata. *Ann. Math. Stud.* **34**, 3–41 (1956)
6. <https://cdn.edureka.co/blog/wp-content/uploads/2017/05/Deep-Neural-Network-What-is-Deep-Learning-Edureka.png>
7. Langley, P.: The cognitive systems paradigm. *Adv. Cogn. Syst.* **1**, 3–13 (2012)
8. Laird, J.E.: *The Soar Cognitive Architecture*. MIT Press, Cambridge (2012)
9. Profanter, S.: Cognitive architectures. Hauptseminar Human-Robot Interaction (2012). <http://profanter.me/static/publications/SeminarCogArch/elaboration.pdf>
10. Laird, J.E., Lebiere, C., Rosenbloom, P.S.: A standard model of the mind: toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Mag.* **38**(7), 13–26 (2017)
11. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. *Psychol. Review* **111**(4), 1036–1060 (2004)
12. Newell, A.: *Unified Theories of Cognition*. Harvard University Press, Cambridge (1990)
13. The Role of Autonomy in DOD Systems. Defense Science Board, July 2012
14. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. *Hum. Factors* **37**(1), 32–64 (1995)
15. Grigsby, S., Crossman, J., Purman, B., Frederiksen, R., Schmorow, D.: Dynamic task sharing within human-UxS teams: computational situation awareness. In: Schmorow, D., Fidopiastis, C. (eds.) *AC 2017, Part II. LNCS*, vol. 10285, pp. 443–460. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_32
16. Schmorow, D., McBride, D.: Augmented cognition (special issue). *Int. J. Hum. Comput. Interact.* **17**(2), 127–130 (2004)
17. Reinerman-Jones, L., Barber, D., Lackey, S., Nicholson, D.: Developing methods for utilizing physiological measures. In: *Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations*. CRC Press, Boca Raton (2010)
18. Matthews, G., Reinerman-Jones, L.E., Barber, D.J., Abich, J.: The psychometrics of mental workload: multiple measures are sensitive but divergent. *Hum. Factors J. Hum. Factors Ergon. Soc.* **57**(1), 125–143 (2015)
19. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. *Acad. Manag. Rev.* **20**, 709–734 (1995)
20. Lyons, J.B.: Being transparent about transparency: a model for human-robot interaction. In: Sofge, D., Kruijff, G.J., Lawless, W.F. (eds.) *Trust and Autonomous Systems: Papers from the AAI Spring Symposium (Technical report SS-13-07)*. AAAI Press, Menlo Park (2013)



Feature Extraction from Social Media Posts for Psychometric Typing of Participants

Charles Li¹, Monte Hancock^{2(✉)}, Ben Bowles³, Olivia Hancock³,
Lesley Perg³, Payton Brown³, Asher Burrell³, Gianella Frank³,
Frankie Stiers³, Shana Marshall³, Gale Mercado³,
Alexis-Walid Ahmed³, Phillip Beckelheimer³,
Samuel Williamson³, and Rodney Wade³

¹ Department of Mathematics and Computer Science,
Mercy College, Dobbs Ferry, USA

² George Herbert Walker School of Business,
Webster University, Webster Groves, USA
practicaldatamining@gmail.com

³ Sirius18, New York, USA

Abstract. Sentiment analysis is an important tool for assessing the dynamic emotional terrain of social media interactions and behaviors [1]. Underlying the shallow emotional phenomenology are deeper and more stable strata, such as culture and psychology.

This work addresses the latter, by applying text mining methods to the assessment of individual psychometrics. A methodology is described for reducing bulk, unstructured text to low-dimensional numeric feature vectors, from which components of the Myers-Briggs Typology Indicator (MBTI) [2] of the text's author can be reliably inferred.

MBTI is a psychometric schema that emerged from the personality theories of Freud and Jung in the early 20th Century, refined and codified by K. C. Briggs and her daughter, I. Briggs-Myers in the 1940's and 50's. This schema positions people along four (nominally independent) axes between pairs of polar motivations/preferences: Extroversion vs. Introversion (E-I); Intuition vs. Sensing (N-S); Feeling vs. Thinking (F-T); and, Judging vs Perceiving (J-P). Under this schema, each person falls into one of 16 psychometric groups, each designated by a four-character string (e.g., INTJ) [3].

Empirical results are shown for text generated during the social media interaction of over 8,600 PersonalityCafe users [4], all of whom are of known MBTI type. Blind tests to validate the features were conducted for a population (balanced by MBTI type), with exemplars based upon text samples having several thousand words each. The feature extraction method presented supports partial (1-letter) MBTI psychometric typing: E-I 95%; J-P 76.25%; F-T 91.25%, N-S 90%. Other results are reported.

Keywords: Social media · Psychometrics · MBTI

1 Introduction

1.1 Moving from Voice Back to Text: A World-Wide Cultural Shift

Human collaborative processes are being revolutionized by the emergence of ubiquitous, completely portable social media. Group decision-making, social interaction, educational instruction, and many other directed and undirected cognitive interactions are now conducted without a single spoken word being exchanged.

As a result of this cultural shift, much social, business, and government communication has moved from postal and telephone exchanges to conversation online. This has fostered the development of world-wide open-source social platforms (e.g., Twitter, Facebook, internet Forums) which daily host billions of text-based interactions. Research is ongoing into questions such as:

- What use can be made of this open-source data?
- What can social media text analysis tell us about its author?
- What the limiting factors for social media text analysis (e.g., how much/what type of data are required? Are results repeatable?)

1.2 Popular Interest in Psychometrics Is Growing

Text mining today generally relies on a combination of statistical and graph theoretic schemas for representing latent information. These schemas are parsed and quantified to obtain information about associations, processes, and conditional probabilities for variables of interest. While some automation exists, the final processing and interpretation is largely manual and ad hoc. The state of the art is described in detail in [6].

The Myers-Briggs Typology Indicator (MBTI) is a psychometric schema that emerged from the personality theories of Freud and Jung in the early 20th Century, refined and codified by K. C. Briggs and her daughter, I. Briggs-Myers in the 1940's and 50's. This schema positions people along four (nominally independent) axes between pairs of polar motivations/preferences: Extroversion vs. Introversion (E-I); Intuition vs. Sensing (N-S); Feeling vs. Thinking (F-T); and, Judging vs Perceiving (J-P). Under this schema, each person falls into one of 16 psychometric groups, each designated by a four character string (e.g., INTJ).

MBTI has become a simple and popular schema for self-analysis among the public. Professional determination of MBTI typology is available; but, informal determination of MBTI typology can be accomplished by taking one (or more) of a large number of free, online, machine-scored multiple-choice tests. A typical example consists of 70 binary-choice check-boxes, and can be completed in less than 10 min. As such, the MBTI is generally a "self-reported" assessment.

Modern applications of MBTI include staffing placement for new employees, corporate team building exercises, student assessment for educational planning, foreign intelligence collection, psychological operations in counter-terrorism, and others.

2 Social Media Data

A Social medium is defined here as any venue supporting public-access pseudo-anonymous self-initiated asynchronous text/image sharing. The term platform will be used here to refer to web-based systems that exist to provide public venues for social interaction and informal information exchange. Preminent examples include Facebook, Twitter, Reddit, Imgur, SnapChat, and many others.

Social media (the plural) have the following attributes:

1. User submissions are referred to variously as:
 - (a) “posts” when the venue is a Forum, Blog (Weblog), or a Chat
 - (b) “messages” when the venue is a messaging system (“tweets” on Twitter)
2. Submissions might or might not be moderated by venue managers, according to guidelines established by the venue owners, and local, state and federal law.
3. Submissions are often organized into threads by topic and/or conversation. A thread is a collection of posts grouped by time, topic, or type.
4. Access to a venue generally requires some kind of “membership” or “registration”.
5. Media often provide various levels of privacy among users (“following, blocking”, etc.).

3 Data Source

The data for this paper were collected through the PersonalityCafe forum in 2017. It consists of snippets (usually whole sentences) of the last 50 posts made by 8,675 people, all of known MBTI type. The entire 59 MB corpus was posted for download by researchers on a web site for Data Science researchers [7].

The average length of the post samples is 1,288 words. All posts are in English. The total size of the corpus is approximately 11.2 million words. Each post sample is tagged with the author’s MBTI type. See Table 2.

The number of words vary from one author to another. Further, the number of posts for the different MBTI types are different. Of special concern is the corpus’ word count imbalance between some of the MBTI types. The “ES” MBTI types have about 20% or fewer of the words of the other types.

Rebalancing the data set by decimating the large classes would have resulted in most of the data being lost. Rather than do this, it was decided to merge the text of each MBTI type to produce 5 exemplars of that type. With 16 MBTI types, this produced 80 MBTI exemplars, the smallest of which would consist of 7,331 words. The table below gives the specifics for each MBTI type. For example (see row 1 of Table 1 below), the corpus contains 190 ENFJ posters. Merging their text in groups of 38 posters produces 5 ENFJ exemplars (henceforth called threads), averaging 51,311 words each. In this way, the 8,645 posters are aggregated into 80 threads, homogeneous by type, each thread consisting of at least 7,331 words, and as many as 399,940 words.

Table 1. Descriptive Statistics for our Corpus for each MBTI type.

MBTI	Code	posts	threads	posts/th	a	b	c	d	e	Total_words	av wrd/th
1	ENFJ	190	1-5	38	50904	47265	54005	52542	51839	256555	51311
2	ENFP	670	6-10	134	177905	177246	181122	170521	178025	884819	176964
3	ENTJ	230	11-15	46	56146	59234	59165	60654	58530	293729	58746
4	ENTP	685	16-20	137	172699	176793	173945	169579	171643	864659	172932
5	ESFJ	40	21-25	8	10187	10373	11988	10415	10162	53125	10625
6	ESFP	45	26-30	9	8872	9923	8595	12632	8882	48904	9781
7	ESTJ	35	31-35	7	9248	9499	8765	9493	7331	44336	8867
8	ESTP	85	36-40	17	22165	17800	19902	23588	20834	104289	20858
9	INFJ	1470	41-45	294	390238	397698	399940	387859	396595	1972330	394466
10	INFP	1830	46-50	366	475687	480938	481103	482092	477432	2397252	479450
11	INTJ	1090	51-55	218	267497	269875	282028	273788	276994	1370182	274036
12	INTP	1300	56-60	260	333145	325943	324394	335942	324376	1643800	328760
13	ISFJ	165	61-65	33	44879	43012	44449	44261	38948	215549	43110
14	ISFP	270	66-70	34	65565	68789	64749	63193	63734	326030	65206
15	ISTJ	205	71-75	41	52768	51390	49986	53629	54100	261873	52375
16	ISTP	335	76-80	67	86332	84960	79579	80068	82656	413595	82719
										<u>11151027</u>	

Table 2. Each post sample is tagged with the author’s MBTI type

Row	A	B	C	D	E	F	G	H	I
1	Row	Thread	post-in-thread	Poster_MBTI_TYPE	Poster_word Count	1	2	3	
2	1	1	1	ENFJ	1260	ABILITY	TO	TRANSFORM	FORM
3	2	1	2	ENFJ	1228	WHAT	ARGUMENTS	THERE	WERE
4	3	1	3	ENFJ	1116	YES	I	HAVE	GONE
5	4	1	4	ENFJ	1128	YEAH	NO	PROBLEM	UMMMM
6	5	1	5	ENFJ	1630	TO	CORRECT	MYSELF	I
7	6	1	6	ENFJ	1657	AHHH	MY	HEART	JUST
8	7	1	7	ENFJ	1354	YOU	ARE	SUCH	A
9	8	1	8	ENFJ	1167	HTTP	//I	IMGUR	COM/EYRHA
10	9	1	9	ENFJ	1003	GOOD	GOOD	THOUGH	I
11	10	1	10	ENFJ	1172	MAKES	ME	FEEL	SPECIAL
12	11	1	11	ENFJ	1389	YOU	KNOW	YOU'RE	AN
13	12	1	12	ENFJ	1788	YOU	FOUND	OUT	SOMEONE
14	13	1	13	ENFJ	1306	WHAT	ARE	YOUR	FEELINGS
15	14	1	14	ENFJ	745	I'M	NOT	INFJ	BUT
16	15	1	15	ENFJ	1292	SORRY	I	KNOW	THIS
17	16	1	16	ENFJ	1203	I	LIKE	TO	FEEL
18	17	1	17	ENFJ	1555	HONESTLY	IN	PERSON	I'VE
19	18	1	18	ENFJ	1444	MOST	OF	THE	ENFJS
20	19	1	19	ENFJ	1268	BUT	BOSS	AIN'T	INFPGIFTNCU
21	20	1	20	ENFJ	1094	I	HAVE	FELT	SAD
22	21	1	21	ENFJ	1466	I	AM	A	LITTLE
23	22	1	22	ENFJ	1457	EXACTLY	MAYBE	IT	HAS

The aggregation process produced a balanced data set having 5 threads for each of the 16 MBTI types.

Because the feature extraction is $O(n^2)$ in the number of threads, aggregation also greatly decreased the processing time required for each experiment.

4 Methodology

Reducing the “bag-of-words” threads to a fixed number of numeric features is most naturally accomplished by histogramming. This places words having some common attribute into bins weighted to create distributions from variable length collections of unstructured text.

4.1 Semantic Mapping: The Category File

To avoid the computational complexity of full-scope computational linguistic (which includes parsing, pronominal reference, stemming, synonymy, etc.), the authors have used approximate semantic tagging based upon a word list having pre-assigned “term categories” and numeric weights. This “Category File” is used for assigning semantic tags (“categories”) and impact scores (“nuances”) to words in threads by a hard match.

The Category File was not created using the MBTI corpus from this work. It was created manually by the authors under a previous effort, using two-years of posts (in colloquial English) from an online Sports Forum. The Forum was first stop-worded to remove “structural terms” (mostly conjunctions, prepositions, articles, etc.); then terms in that 10-million word corpus occurring with frequency above a significance threshold were collected. The resulting set consisted of 4,116 terms. These words were collected into 126 subjectively defined “semantic categories”, and each word assigned a “nuance”, subjectively expressing its “impactfulness” as an integer in the range -3 to $+3$. Negative and positive values indicate “bad” and “good” impact, resp. See Table 3.

Terms in a thread that do not occur in the category file are discarded; terms in a thread that are found in the category file are accumulated into the corresponding histogram bin for that thread.

The category file has 4,116 entries (rows), with five values in each row: Term, Part-of-Speech, Term Category, Frequency in Reference Corpus, Nuance. See Table 3.

4.2 Computing the Components of Thread Similarity

The extraction of features must preserve the salient similarities and differences between the entities being analyzed. Typically, the encoding of any discriminating information across the data set is unknown. It is, therefore, customary to select multiple entity attributes, encode them as features, and perform empirical experiments to determine their discriminating power.

The following term statistics were computed for each term in each thread:

- Term Category (CAT)
- Term Nuance
- Term Frequency (TF)
- Term Inverse Document Frequency (idf)
- Term Frequency times Inverse Document Frequency (Tf.idf)

Each term in a thread will, in general, have different values for these components.

Given that there are thousands of terms in a thread, unifying these components directly (e.g., by placing them into an ordered n-tuple) would produce large,

Table 3. Category File entries give word attributes

Term	POS	Category	FREQUENCY	Nuance	
HEAVENLY	ADJ		1	29	3
AWESOME	ADJ		1	25	3
FANTASTIC	ADJ		1	24	3
GOLDEN	ADJ		1	23	3
HOLY	ADJ		1	23	3
HUGE	ADJ		1	23	3
BLESSED	ADJ		1	22	3
OUTSTANDING	ADJ		1	22	3
AMAZING	ADJ		1	16	3
BEAUTIFUL	ADJ		1	16	3
SUPER	ADJ		1	16	3
INCREDIBLE	ADJ		1	14	3
PERFECT	ADJ		1	12	3
HELLUVA	ADJ		1	11	3
SPIRITUAL	ADJ		1	8	3
STELLAR	ADJ		1	8	3
THRILLED	ADJ		1	8	3
BRILLIANT	ADJ		1	7	3
EPIC	ADI		1	7	3

unsynchronized vectors having different dimensions. Such a representation requires conformation of some kind to make it suitable for thread analysis. Histogramming mitigates this problem.

5 Determining Term Category and Nuance for Each Term in Each Thread

These are determined by hard match to the corresponding term entry in the Category File.

The categories provide a coarse semantic mapping. The 126 categories in the category file include “Salient Adjectives”, “Salient Adverbs”, “Parts of the Body”, “Filial Relationships”, “Major Cities and States”, “Common Names”, etc. The rationale is that pairs of threads using terms in related categories are likely to be addressing the same topic, making them more similar. Whether this is true was investigated by experiment, and is discussed below.

6 Computing Term Frequency, Tf (Document, Term), for Each Word in Each Thread

First, compute T1 for each word in each thread. T1(thread, "word") is the number of times, counting multiplicities, that "word" occurs in the thread under consideration:

$$T1(\text{thread}, \text{"word"}) = \# \text{ occurrences of "word" in document, counting multiplicity.}$$

The T1 value of the most frequently occurring word in a document is called T2 (thread) for that thread, and is used for normalization:

$$T2(\text{thread}) = \# \text{ occurrences of most frequently occurring word in the thread}$$

We define Term Frequency each term in each thread as:

$$Tf(\text{document}, \text{term}) = T1(\text{document}, \text{term}) / T2(\text{document})$$

7 Computing Inverse Document Frequency, Idf, for Each Word in a Document

Let I1 be the number of threads in the entire corpus:

$$I1 = \# \text{ threads in the corpus}$$

Then, for each term in the entire corpus, count the number of threads in the entire corpus that contain that term; call this

$$I2(\text{"word"}) = \# \text{ threads in corpus containing "word"}$$

Form the expression: $idf(\text{"word"}) = \log(I1 / (1 + I2(\text{"word"})))$

Notice that the argument of the log is essentially the reciprocal of $I2/I1$, a “document frequency”, so idf is called the inverse document frequency. The “1” in the denominator prevents division by zero, and the log pulls the result into a reasonable range, since this ratio is usually very large. It is customary to use either the $\ln(x)$, or $\log_2(x)$. The “1” in the denominator is not used by all researchers; this matters little when the number of words is large, as it is in most applications.

8 Computing the Tf.Idf Score for Each Word in Each Thread

Note: Tf.idf (usually vocalized as a 5-letter acronym: “tfidf”) is the standard term for the product (Tf)(idf)

From Tf and idf, we obtain the Tf.idf score for each term in each thread by multiplying Tf (“thread”, “word”) by idf(“word”):

$$\text{Tf.idf}(\text{"thread"}, \text{"word"}) = (\text{Tf}(\text{"thread"}, \text{"word"})) \times (\text{idf}(\text{"word"}))$$

The idea behind the Tf.idf score is that words that occur often in a particular thread, but are relatively rare in other thread, probably have meanings strongly related to the “meaning” of the threads in which they occur often: semantically significant terms in a thread are naturally expected to have high Tf, and high idf in that thread.

9 Developing Weighted Histograms from Which to Extract Features

Each thread can be viewed as a distribution by its word-frequency histogram using the terms in the Category File. This histogram will have 4,116 bins, because the Category File has entries for 4,116 terms.

The unweighted normalized frequency histograms for each MBTI type in the corpus are shown in Fig. 1. Notice the absence of artifacts that might indicate problems arising from disparity in word count among the MBTI types.

These unweighted normalized frequency histograms give the word selection distribution for each MBTI type. To fuse information from other term statistics for analysis, the corresponding terms bins can be weighted by those statistics. For example, if the Nuance scores for each term are multiplied by the corresponding BIN count in the histogram, we obtain a Nuance weighted histogram. This weighting can be carried out with any term statistic. The statistic will have more impact where the BIN count for that term is highest, automatically weighting the statistic by its “prevalence”.

In particular, the term statistics previously described can be used to scale the word-frequency histograms for threads. Thread similarity is performed by viewing the various weighted, term-frequency histograms as vectors in a 4,116 Euclidean space, where many metrics are available.

10 Creating Weighted Histograms for Threads

To create histograms for extracting features for a thread, the following are collected for each word it contains:

- CAT: binary Category (“1” if term is in Category File, else “0”)
- Nuance: the nuance value for the term in range [− 3, +3]

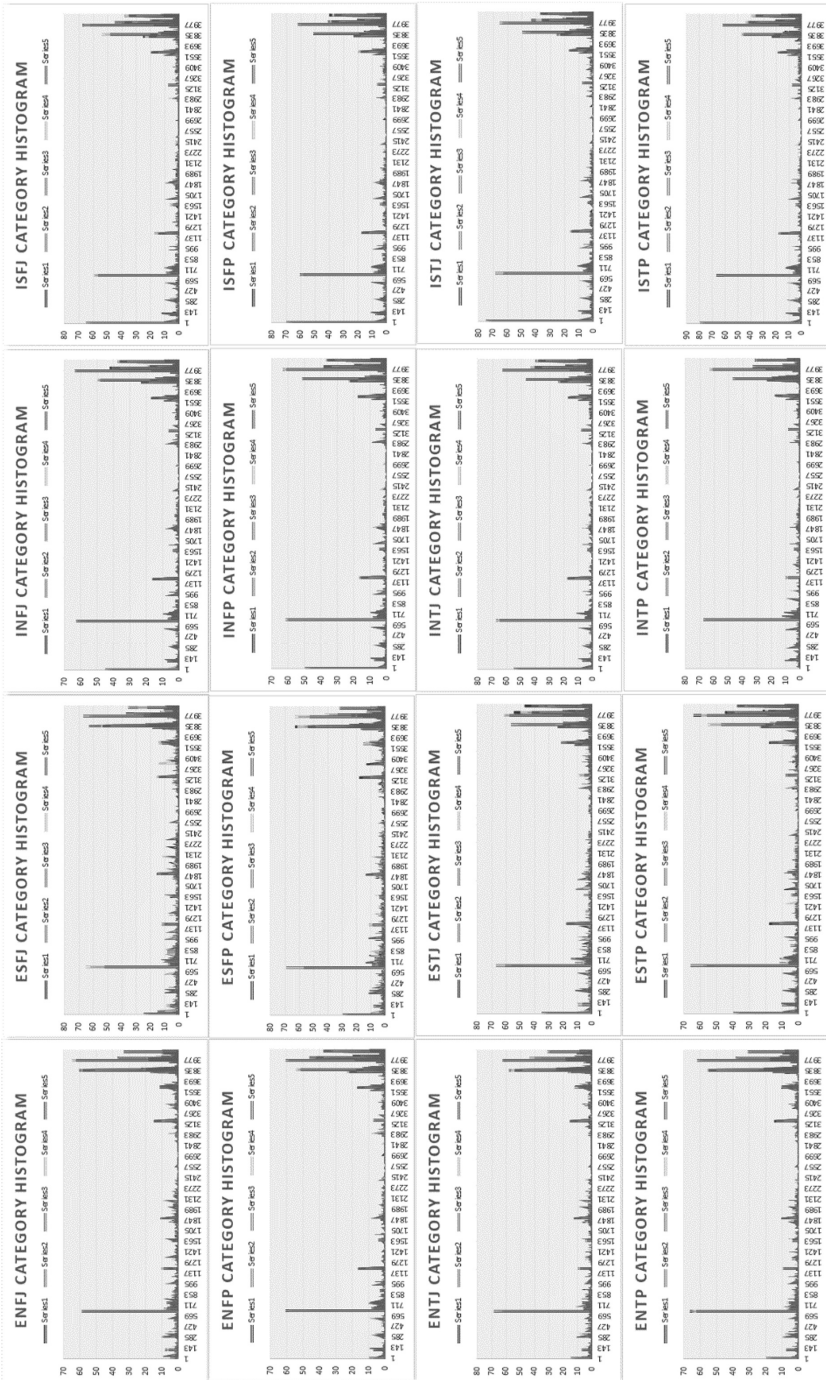


Fig. 1. The unweighted normalized frequency histograms for each MBTI type in the corpus.

- Term Frequency: (Tf) for the term in the document
- Inverse Document Frequency: (idf) for the term in the document
- TF.idf: the product of Tf and idf values for the term

Define the Relative Frequency for a term row as: $nu = T1/I2$, and accumulate four weighted histograms as follows:

```
CAT(author,tagrows)=nu*Category    (weighted "0" or "1")
NIV(author,tagrows)=nu*nuance      freq.-weighted nuance
TFV(author,tagrows)=nu*Tf          freq.-weighted Tf
IDFV(author,tagrows)=nu*idf        freq.-weighted idf
TFIDFV(author,tagrows)=nu*Tfidf    freq.-weighted Tfidf
```

Inferring Inter-Author Distances from Histograms Twenty-six metric values are then computed from the four histograms above. They fall into six groups, depending upon which of histograms are used:

- take dot products of histograms
- use ArcCos turns dot products into angles between normalized histograms
 - $[-\pi, \pi]$ rescaled to $[-1, 1]$, so 0 implies perpendicularity
- compute RMS vector distance between histograms as 4,116-vectors
- compute Euclidean distance between histograms as 4,116-vectors

The computation of the feature values is performed by five software routines: Faze_A, Faze_B, Faze_C, Faze_D, and Faze_Alpha. Faze_Alpha is the controlling routine, calling the others in order A through D.

This produces twenty-seven sets of distance matrices between the 80 threads described in the next section, Feature Extraction Detailed Flow.

11 Feature Extraction Detailed Flow

Faze_Alpha: Sequentially invoke Fazes A - E.
Faze_A: Read operating parameters from disc: Faze.prm
Faze_A: Read semantic tags from disc: tags.csv
Faze_A: Read in the entire text corpus from disc
Faze_A: Create an abridged Text Corpus (aTC): AnnText.csv
Faze_A: aTC all caps, no punctuation, comma delimited
Faze_A: aTC keeps only tagfile words (no misspellings)
Faze_A: Create word histogram for aTC: taghisto.csv
Faze_B: Read operating parameters from disc: Faze.prm
Faze_B: Read wordlist: wordlist.csv
Faze_B: Read the tagfile from disc: tags.csv
Faze_B: Scan wordlist threads: wordlist.csv
Faze_B: Compute term scores for all words in all threads:
Faze_B: T1 = occurrences of term in document
Faze_B: T2 = occurrences most frequently occurring term
Faze_B: $Tf(\text{term}, \text{document}) = T1/T2$
Faze_B: I1 = # documents in the corpus
Faze_B: I2 = # documents in corpus containing term
Faze_B: $idf(\text{term}, \text{corpus}) = \log(I1/(1+I2))$
Faze_B: $Tf.idf = Tf(\text{term}, \text{thread}) * idf(\text{term}, \text{corpus})$
Faze_B: For each word in each thread, tabulate:
Faze_B: Thread, Category, Nuance, T1 , T2 , I1 , I2,
Tf, idf, Tf.id: stored in file Tfidf.csv
Faze_C: Read operating parameters from disc: Faze.prm
Faze_C: Read in the thread term scores: Tfidf.csv
Faze_C: Build the thread vector for this thread by concatenating records holding the term scores for each of the terms in this thread (even when the thread is spread across rows [and so, multiple same-MBTI-type authors] in the original text corpus).
Faze_C: Write out the resulting thread vector, appending records term-by-term, so that a single output row has the various term scores for all the terms in this thread in tag file order. This thread vector will be a weighted roll-up of all the term scores for this thread; it is similar to a relative frequency histogram for this thread, and will have 4116 bins.

NOTE: histogram entries are freq.-novelty weighted by multiplication by the BIN T1/T2.

By this process, each thread generates five frequency histograms of 4116 BINS:

Cat(4116), histogram of term category counts
 Nu(4116), histogram of term Nuances
 TF(4116), histogram of Term Frequencies
 IDF(4116), histogram of Inverse Document Frequencies
 TFIDF(4116), histogram of term TF.idf Scores

Each histogram gives a different information-theoretic view of a thread.

These five histograms are then combined in various ways to obtain hybrid, non-linear distance ‘indicators’ of relatively low dimension. These combination strategies included: Euclidean distance, RMS distance, angle between histograms, and sum-of-squares.

These were collected into six groups, based upon which underlying statistic was used.

For each pair of threads, th1 and th2, twenty-seven thread similarity scores (“Metrics”) were created, as follows:

Group 1 thread similarity scores: Category

Metric 1 MIX_SOQ(th1,th2): Non-Linear Mixing Feature
 Metric 2 CAT_Edist(th1,th2): Euclidean distance between
 unweighted term histograms
 Metric 3 CAT_RMS(th1,th2): RMS distance between
 unweighted term histograms
 Metric 4 CAT_SOQ(th1,th2): sum-of-squares of CAT
 Edist, RMS, CAT_CAT
 Metric 5 CAT_CAT(th1,th2): Cosine Distance between
 unweighted term histograms

Group 2 thread similarity scores: Nuance

Metric 6 NIV_Edist(th1,th2): Euclidean distance between
 Nuance-weighted term histos
 Metric 7 NIV_RMS(th1,th2): RMS distance between
 Nuance-weighted term histos
 Metric 8 NIV_SOQ(th1,th2): sum-of-squares of NIV
 Edist, RMS, NIV_NIV
 Metric 9 NIV_NIV(th1,th2): Cosine distance between
 Nuance-weighted term histos

Group 3 thread similarity scores: Term Frequency

Metric 10 TF_Edist(th1,th2): Euclidean distance between
Tf-weighted term histos

Metric 11 TF_RMS(th1,th2): RMS-distance between
Tf-weighted term histos

Metric 12 TF_SOQ(th1,th2): sum-of-squares of TF
Edist, RMS, Tf_Tf

Metric 13 TF_TF(th1,th2): Cosine distance between
Tf-weighted term histos

Group 4 thread similarity scores: Inverse Document Frequency

Metric 14 IDF_Edist(th1,th2): Euclidean distance between
Idf-weighted term histos

Metric 15 IDF_RMS(th1,th2): RMS distance between
Idf-weighted term histos

Metric 16 IDF_SOQ(th1,th2): sum-of-squares of idf
Edist, RMS, idf_idf

Metric 17 IDF_IDF(th1,th2): Cosine distance between
idf-weighted term histos

Group 5 thread similarity scores: TF.idf

Metric 18 TFIDF_Edist(th1,th2): Euclidean distance between
TF.idf-weighted term histos

Metric 19 TFIDF_RMS(th1,th2): RMS distance between
Tf.idf-weighted term histos

Metric 20 TFIDF_SOQ(th1,th2): sum-of-squares of Tf.idf
Edist, RMS, Tf.idf_Tf.idf

Metric 21 TFIDF_TFIDF(th1,th2): Cosine distance between
Tf.idf-weighted term histos

Group 6 thread similarity scores: Histogram Dot products

Metric 22 NIV_TF_SOQ(th1,th2): sum-of-squares of
(NIV_SOQ)*(TF_SOQ)

Metric 23 NIV_IDF_SOQ(th1,th2): sum-of-squares of
(NIV_SOQ)*(IDF_SOQ)

Metric 24 NIV_TFIDF_SOQ(th1,th2): sum-of-squares of
(NIV_SOQ)*(TFIDF_SOQ)

Metric 25 TF_IDF_SOQ(th1,th2): sum-of-squares of
(TF_SOQ)*(IDF_SOQ)

Metric 26 TF_TFIDF_SOQ(th1,th2): sum-of-squares of
(TF_SOQ)*(TFIDF_SOQ)

Metric 27 IDF_TFIDF_SOQ(th1,th2): sum-of-squares of
(IDF_SOQ)*(TFIDF_SOQ)

The pairwise thread similarities (“Metrics”) for the 80 threads are placed into twenty-seven 80-by-80 pairwise distance matrices. For example, the 3,160 (non-trivial) RMS distances in inverse document frequency between threads th1 and th2 are in IDF_RMS(th1,th2), Table 4.

Table 4. Distances between pairs of threads

doc1	doc2	CATdotCA	CAT_CAT	NIV_Edist	NIV_RMS	NIVdotNI	NIV_NIV	TF_Edist	TF_RM!
1	2	3.86628	0.449326	0.007004	0.509387	0.239858	0.007738	0.000121	0.2572
1	3	3.477523	0.325508	0.005074	0.395174	0.224012	0.007487	0.000117	0.2611
1	4	3.177049	0.317305	0.004946	0.388308	0.223777	0.006502	0.000101	0.2388
1	5	3.159429	0.350005	0.005456	0.415898	0.224584	0.006653	0.000104	0.2348
1	6	6.942426	1.346483	0.020988	1.359286	0.184939	0.024109	0.000376	0.2058
1	7	6.002892	1.156543	0.018027	1.171342	0.184732	0.024429	0.000381	0.2069
1	8	6.297103	1.223485	0.01907	1.237458	0.184456	0.023707	0.00037	0.2016
1	9	7.595522	1.403152	0.021871	1.417944	0.202106	0.026327	0.00041	0.215

Features for clustering and supervised learning are inferred from these distance matrices [8]. This is done by hypothesizing the existence of a low-dimensional point set having the same distance matrix as that developed for the pairs of threads. Such a point set is computed using gradient descent, and the coordinates of these points become abstract features for the corresponding thread.

Each row shows the twenty-seven different distances that have been computed between the pair of threads specified in the two leftmost columns. Twenty-seven, 6-dimensional feature sets were developed for the corpus in this way. Processing time is about 25 min for the 80-thread case described here.

```
Faze_D: Read operating parameters from disc: Faze.prm
Faze_D: Read in the 27 distance matrices, all held in a
single file: Delta_File.csv
Faze_D: For each distance matrix, use gradient descent to
infer low-dimensional Torgerson Coordinates for each
thread.
Faze_D: Save these 27 feature sets on disc for evalua-
tion.
```

Here are the 27 feature sets created for the corpus:

Group 1: Semantic Categories

- CATdotCAT inter-thread dot product of category
- CAT_CAT inter-thread Cosine dist. using category

Group 2: Term Nuances

NIV_Edist inter-thread Euclidean distance using Nuances
 NIV_RMS inter-thread RMS distance using Nuances
 NIVdotNIV inter-thread dot product of Nuances
 NIV_NIV inter-thread Cosine dist. using Nuances

Group 3: Term_Frequencies

TFV_Edist inter-thread Euclidean distance using Tf
 TFV_RMS inter-thread RMS distance using Tf
 TFVdotTFV inter-thread dot product of Tf
 TFV_TFV inter-thread Cosine dist. using Tf

Group 4: Inverse Document Frequencies

IDFV_Edist inter-thread Euclidean distance using idf
 IDFV_RMS inter-thread RMS distance using idf
 IDFVdotIDFV inter-thread dot product of idf
 IDFV_IDFV inter-thread Cosine dist. using idf

Group 5: Tfidf

TFIDFV_Edist inter-thread Euclidean distance using Tf.idf
 TFIDFV_RMS inter-thread RMS distance using Tf.idf
 TFIDFVdotTFIDFV inter-thread dot product of Tf.idf
 TFIDFV_TFIDFV inter-thread Cosine dist. using Tf.idf

Group 6: Cross_Dots

NIVdotTFV inter-thread dot product of Nuance & Tf
 NIVdotIDFV inter-thread dot product of Nuance & idf
 NIVdotTFIDFV inter-thread dot product of Nuance & Tf.ID
 TFVdotIDFV inter-thread dot product of Tf & idf
 TFVdotTFIDFV inter-thread dot product of Tf & Tf.idf
 IDFVdotTFIDFV inter-thread dot product of idf & Tf.idf

12 Feature Validation Experiment

To determine whether feature sets derived by the method described here captured information distinguishing threads (and so, their authors) by MBTI type, each feature set was subjected to leave-one-out validation using a simple nearest neighbor classifier. See Table 5 (Fig. 2).

Table 5. The highest and lowest accuracy confusion matrices (showing precision, recall, and classification accuracy) for the single-letter MBTI classes.

<p>CAT_Edist is the Euclidean distance based upon weighted nuances</p>		<p>BEST EI</p>		<p>Distance metric 2</p>	
Act\Est					
E: Extrovert		I: Introvert	PRECISION	Closest Class Accuracy:	90%
I: Introvert	36	4	90%	Class 1 consists of	40 vectors (50%)
	4	36	90%	Class 2 consists of	40 vectors (50%)
<p>TFIDF_RMS is the RMS distance using weighted Tf,idf</p>		<p>WORST EI</p>		<p>Distance metric 19</p>	
Act\Est					
E: Extrovert		I: Introvert	PRECISION	Closest Class Accuracy:	35%
I: Introvert	14	26	0.35	Class 1 consists of	40 vectors (50%)
	26	14	0.35	Class 2 consists of	40 vectors (50%)
<p>TF_IDF_SOQ is the sum-of-squares inTer-product of weighted Tf and idf</p>		<p>BEST FT</p>		<p>Distance metric 25</p>	
Act\Est					
F: Feeling		T: Thinking	PRECISION	Closest Class Accuracy:	91.25%
T: Thinking	35	5	0.875	Class 1 consists of	40 vectors (50%)
	2	38	0.95	Class 2 consists of	40 vectors (50%)
<p>CAT_RMS is the RMS distance based upon weighted nuances</p>		<p>WORST FT</p>		<p>Distance metric 3</p>	
Act\Est					
F: Feeling		T: Thinking	PRECISION	Closest Class Accuracy:	37.5%
T: Thinking	15	25	0.375	Class 1 consists of	40 vectors (50%)
	25	15	0.375	Class 2 consists of	40 vectors (50%)
<p>MIX_SOQ is the Non-Linear Mixing Feature</p>		<p>BEST JP</p>		<p>Distance metric 1</p>	
Act\Est					
J: Judging		P: Perceivi	PRECISION	Closest Class Accuracy:	76.25%
P: Perceiving	32	8	0.8	Class 1 consists of	40 vectors (50%)
	11	29	0.725	Class 2 consists of	40 vectors (50%)
<p>IDF_RMS is the RMS distance using weighted idf</p>		<p>WORST JP</p>		<p>Distance metric 15</p>	
Act\Est					
J: Judging		P: Perceivi	PRECISION	Closest Class Accuracy:	40%
P: Perceiving	22	18	0.55	Class 1 consists of	40 vectors (50%)
	30	10	0.25	Class 2 consists of	40 vectors (50%)
<p>CAT_SOQ is the sum-of-squares for CAT</p>		<p>BEST NS</p>		<p>Distance metric 4</p>	
Act\Est					
N: Intuitive		S: Sensing	PRECISION	Closest Class Accuracy:	90%
S: Sensing	36	4	0.9	Class 1 consists of	40 vectors (50%)
	4	36	0.9	Class 2 consists of	40 vectors (50%)
<p>IDF_RMS is the RMS distance using weighted idf</p>		<p>WORST NS</p>		<p>Distance metric 15</p>	
Act\Est					
N: Intuitive		S: Sensing	PRECISION	Closest Class Accuracy:	43.75%
S: Sensing	12	28	0.3	Class 1 consists of	40 vectors (50%)
	17	23	0.575	Class 2 consists of	40 vectors (50%)

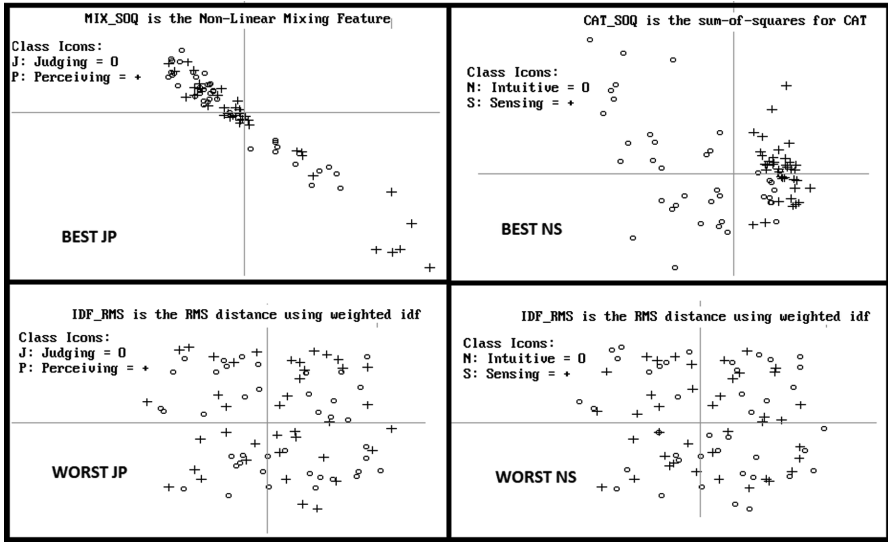


Fig. 2. Here are some typical pair-plots projected on the first two of 6 features generated.

13 Conclusions

The “metric numbers” here refer to those defined in Sect. 11 above.

The most successful metrics for distinguishing extroversion/introversion (E/I) are metrics 1, 2, 4 and 5. The recall for extroverts range from 80% to 95% among these metrics, while the recall for introverts range from 82.5% to 90%. The precision for extroverts range from 84% to 90% among these metrics, while the precision for introverts range from 81.8% to 94.3%. The other metrics were not as successful in distinguishing E/I and the recall and precision figures from most of those metrics were typically in the 40% to 70% range.

Metrics 1 and 2 compare the category bits of one class to another, while metrics 4 and 5 compare their weighted nuances. The data suggests that there is a clear distinction between the words favored by the extroverts and those favored by the introverts. The category bits involve sorting words into various categories based on their meanings. The strong results from the metrics involving category bits could possibly mean the extroverts in the data set were gravitated toward discussing certain subjects and the introverts toward other subjects, or perhaps extroverts and introverts write/speak differently. The results from metrics 4 and 5, which involve comparing weight nuances, would also follow as a result if this supposition were true.

While some of the other metrics also involve the weighted nuances of the words, these metrics typically compare the weighted nuances with some other factor that is stripped of the semantics, stripped of the meanings of the words. For example, the least successful metric for distinguishing E/I is metric 19, which compares the weighted

nuances for one class against the term frequencies of words for the other class. The recall and precision figures for both extroverts and introverts were all 35% using this metric.

The most successful metric for distinguishing feeling/thinking (F/T) is metric 4. The recall for feeling is 87.5% while the recall for thinking is 95%. The precision for feeling is 94.6% while the precision for thinking is 88.4%. Note that metric 4 involves the weighted nuances of words, so this result is not surprising.

Metrics 1, 2, 5, 11, 14, 15, 18, 19 and 26 have recall and precision figures for F/T typically in the 70% to 80% range. The other unsuccessful metrics have recall and precision figures typically in the 40% to 70% range. Metrics 1, 2 and 5 emphasize the effect of the semantic content of words and thus were expected to perform respectably in distinguishing F/T. Metrics 11, 14, 15, 18, 19 and 26 involve non-semantic factors such as the inverse document frequency and term frequency of words. These other factors show up in both the semi-successful as well as the unsuccessful metrics for distinguishing F/T, so the importance of these factors is not without question.

The most successful metrics for distinguishing judging/perceiving (J/P) are metrics 1 and 2. The recall and precision figures for judging and perceiving for both of these metrics range from 70% to 80%. The recall and precision figures for the less successful metrics were mostly in the 50% to 70% range, with quite a few in the 60% to 70% range.

These results suggest that while the categories of words used in metrics 1 and 2 was the most successful factor in distinguishing J/P, many of the other metrics were not too far behind. Furthermore, success in the 70% to 80% range when it comes to recall and precision is not extremely remarkable in the first place. These results possibly suggest that J/P is not something that is strongly distinguishable from the words that people use, but rather from their actions or behaviors instead. Note that the MBTI test itself asks questions related to actions and behaviors when assessing J/P.

The most successful metrics for distinguishing intuitive/sensing (N/S) are metrics 1, 2, 4, 5, 20 and 23. The recall and precision figures for N/S typically range from 80% to 90% for these metrics, although some of the figures were as high as 97.5%, such as the case of the recall for intuitive for metric 5. The less successful metrics had recall and precision figures typically in the 50% to 70% range. The success of metrics 1, 2, 4 and 5 in distinguishing N/S suggests that people who classify as intuitives write/speak differently from those who classify as sensing, or perhaps they gravitate toward different subjects for discussion.

Of particular note is metric 23, which compares the term frequency of words in one class to the tf.idf scores of another. This is a non-semantic factor and the recall and precision figures for N/S ranged from 82.5% to 95% for metric 23.

Overall, metrics 1 and 2 were the most successful in distinguishing all four of the pairs E/I, F/T, J/P and N/S. This suggests the various pairs can potentially be distinguished by the way people write/speak, or perhaps by the subjects they are interested in discussing. J/P was the hardest pair to distinguish since it involves looking at people's actions and behaviors to distinguish, not just words.

14 Future Work

Some potential directions to go for future work include analyzing combinations of the various Myers-Briggs binary classifications, such as NT or SJ. There is literature to suggest that such pairings are significant. For example, people who type as NTs are referred to as “rationals” in Keirsey’s temperament sorter, and those who type as SJ are called “guardians”.

Another possible direction for future work is to study the verbosity of people as it relates to their Myers-Briggs personality types. For example, it is plausible that introverts are more verbose in writing than extroverts, given the introverts’ preference for writing over speaking. This direction might shine light on the role played by the non-semantic factors studied in this project such as term frequency, and this direction might greater fine-tune the metrics used to distinguish personality types.

Acknowledgements. We gratefully acknowledge the support of the Sirius Project, and the INTJForum [5].

References

1. <https://www.thebalance.com/what-is-social-sentiment-and-why-is-it-important-3960082>
2. <http://www.myersbriggs.org>
3. <http://www.humanmetrics.com>
4. <http://PersonalityCafe.com>
5. <http://IntjForum.com/>
6. Lee, S., Song, J., Kim, Y.: An empirical comparison of four text mining methods, 2010Utility. In: 18th International Conference on Human Computer Interaction, Toronto, Canada, July 2016
7. <https://www.kaggle.com>
8. Lebanon, G.: Information geometry, the embedding principle, and document classification. In: Proceedings of the 2nd International Symposium on Information Geometry and its Applications, pp. 101–108 (2005)

Bibliography

9. Hancock, M., Sessions, C., Lo, C., Rajwani, S., Kresses, E., Bleasdale, C., Strohschein, D.: Stability of a type of cross-cultural emotion modeling in social media. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) AC 2015. LNCS (LNAI), vol. 9183, pp. 410–417. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20816-9_39
10. Hancock, M.: Practical Data Mining. CRC Press, Boca Raton (2011)
11. Delmater, R., Hancock, M.: Data Mining Explained: A Manager’s Guide to Customer-Centric Business Intelligence. Digital Press, Boston (2001)

12. Hancock, M., et al.: Field-theoretic modeling method for emotional context in social media: theory and case study. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) AC 2015. LNCS (LNAI), vol. 9183, pp. 418–425. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20816-9_40
13. Carapinha, F., et al.: Modeling of social media behaviors using only account metadata. In: Schmorow, D.D., Fidopiastis, C.M. (eds.) AC 2016. LNCS (LNAI), vol. 9744, pp. 393–401. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39952-2_38



Intermediate Information Grouping in Cluster Recognition

Chloe Chun-wing Lo¹(✉), Markus Hollander², Freda Wan³,
Alexis-Walid Ahmed³, Nikki Bernobić³, Nick Nuon³,
and Michael Shrider³

¹ Cherrypicks Limited, Kowloon, Hong Kong
chloelcw@cherrypicks.com

² Center for Bioinformatics, Saarland University, Saarbrücken, Germany

³ Sirius18, New York, USA

Abstract. Unsupervised clustering takes human brains a split second to complete in 2D space, while existing algorithms require many iterations involving all data points and an initial number of clusters ‘k’ to provide meaningful results. This initial ‘k’ cannot be provided by human if the data is in higher dimension where visualization is practically impossible. Attempts to calculate this value have low performance and give ambiguous results which are unhelpful to human judgment. This presents great motivation to search for a method to provide that initial ‘k’ in higher dimension. Human brains naturally group things in proximity together. By imitating this process and creating a middle process of grouping the data into subregions and mapping the data in each region into a bitmap of data densities, estimating the centroid locations and number of clusters can be simplified into a process of local maxima detection. A run of the algorithm on 2D data proved that it was effective for data with Gaussian-like distribution with some tolerance to overlapping. The algorithm therefore has great potential for data of higher dimension after generalization. This algorithm gives unambiguous initial ‘k’ and fairly accurate estimation of centroids which can speed up various popular clustering algorithms, including the k-means and Gaussian mixture models. Future research on middle grouping processes in human cognition, which may prove valuable in providing better machine learning algorithms, are also called for.

Keywords: Clustering · Proximity grouping · Centroid estimation

1 Literature Review

Clustering is one of the basic data science tasks, where the goal is to separate unlabeled datasets into groups so that data points within a group have more in common in between them than with data points outside of it. This definition highlights typical non-trivial problems a good clustering algorithm must be able to handle, for example:

- What is a cluster?
- How many clusters are there?

- What is the definition of “similarity”?
- Where does a cluster end?
- What shape can a cluster have?

Such challenges have led to a very wide variety of approaches [1], making a comprehensive summary difficult to establish, but most obvious families of clustering algorithms are defined according to the metric/method they use to evaluate differences between points [2, 3]. A reminder of the most commonly seen unsupervised algorithms is provided:

- **Distribution based clustering methods** start by assuming an underlying statistical distribution and will then try to maximize the fit to the data by tuning parameters of the generating function; the initialization of such algorithms is a critical phase that sometimes hinders repeatability as they are prone to getting stuck in local minima, and the choice of the distribution is often a gamble. One of the most successful examples would be Gaussian Mixture Models.
- **Hierarchical clustering methods** attempt to either start from the assumption that all points are in the same big cluster then separate them progressively into a constellation of point-like clusters while keeping track of the order in which subgroups secede; alternatively, they may start from the bottom and successively merge all clusters into one while keeping track of the assimilations; either way, hierarchical clustering is based on a measure of distance (generally Euclidean or Manhattan) between points and are thus sensitive to outliers with an inability to correct themselves during their execution.
- **Centroid-based clustering methods**, largely based on one of the most widely-used clustering algorithm, the K-means [4], attempt to find clusters of data by calculating a centroid which is not necessarily part of that dataset; the choice of the way centroids are arranged is determined by a cost function, however the choice of the number of centroids itself is still a largely unresolved question, often requiring user input which can be absent or unreliable in high dimensional and/or messy data, or using methods adapted for narrow cases. Since that family needs an a priori number of clusters to start, dozens of techniques have been devised to address that blind spot, called cluster validation techniques that are generally broken down into three different kinds: internal, external, and relative [5]. Internal methods describe measures that depend only on the data used in the clustering itself and the clustering schema, while external methods rely on additional information, for example, a user specific intuition. Relative methods consist in running different clustering schemas with different parameters then selecting the «best» one according to a validity index or predefined criterion. One such popular method with the K-mean is the elbow method, where the clustering is done with different K then some measure of error applied. When adding another cluster K_{n+1} stops improving performance, K_n is found to be optimal. All these methods have in common their computational cost, and the need to use a few of them in parallel to offset their respective biases and assumptions, leading to the circulation of «recipes» of these validation techniques with various popularities.

- **Density based clustering methods** [6] define clusters by a measure of the local density in data points, with the assumption that boundaries will tend to have a low density largely due to noise; such methods thus handle outliers and noise quite well since they generally cannot compete in quantity and density with more plausible clusters in a given representative dataset; the major difficulty in those methods is that a predefined reach is often used to prevent overfitting, so that flexibility of the algorithm with regard to the scales it can simultaneously resolve is limited, and detecting boundaries relies on detecting a drop in density, which can be difficult to ascertain and is very dependent on the consistency across clusters. At the frontiers with Topological Data Analysis and Artificial Neural Networks, approaches such as t-SNE [7] or self-organizing maps (SOM or Kohonen Maps [8]) have been used because they generally allow for topology preserving dimensionality reduction of datasets while making even high-dimensional clusters more apparent to the naked eye; such methods can however be thought of as extended versions of previously described algorithms, SOMs in particular have been shown to be rigorously identical in their late steps to K-means [9].
- **Bioinspired algorithms**, among which herd algorithms are particularly promising [10], characterized by a tendency for decentralized processing relying on emergent clustering, have been steadily developed since the classical ant colony inspired algorithm was first laid out; they are robust to noise and dynamic, making them suitable for time-dependent data analysis [11], whereas previous techniques are generally confined to static data (although recent developments are starting to change that [12]).

However, none of the above seems to truly parallel with the performance of human brain. Humans are able to detect cluster of objects by proximity grouping [13] in split second. Since data points in close proximity suggest a denser data points relative to their surroundings, it is of great interest whether a local density maxima could successfully detect the number of clusters with a better efficiency and performance. An algorithm that first groups the data into sub regions and comparing densities in each region considering only their neighboring regions is then developed in the hope of providing a more accurate and reliable estimation of the number of clusters in a given dataset, and also possibly the estimation of the location of their centroids.

2 Procedure

We tested our algorithm on all datasets from <https://cs.joensuu.fi/sipu/datasets/>. Here, we are showcasing the following 4 datasets: S-sets S-3 and S-4, A-set A-3 and Birchset Birch 1.

2.1 Hardware and Software

The algorithm is run on a MacBook Pro (Retina, 13-in., late 2013), Intel Core Duo i5, 8 GB 1600 MHz DDR3 RAM, macOS High Sierra version 10.13.2 using Python3 v3.6.3 with Seaborn v0.8.1, Pandas v0.21.0 and NumPy v1.13.3.

2.2 Algorithm

The algorithm consists of two parts. First, it searches for the best grid size to divide the dataset into subregions, and then scans for local maxima.

Determination of Optimal Grid Size. In the following passages, the letter m denotes how many portions each side of the bounding box of the data is divided into (which results in m^2 grid boxes in total). The computation of the best grid size is done as follows:

1. Set the initial m is set to be the box length divided by 2.
2. The bounding box of the data is segmented into m^2 grid boxes.
3. The number of data points present in each grid box is calculated.
4. Check whether the number of boxes with single data is less than the tolerance level (set at 0.5% in this paper) of all data points.
5. If the above criteria is not met, m will be decreased exponentially by multiplying the adjustment factor (set at 0.0975 in this paper).
6. The above process repeats until the criteria in 4 is met.

Identification of Local Maxima. The local maxima are identified as follows:

1. Go through all the grid boxes from the top-left to bottom-right by a 3-by-3 moving window.
2. Mark the coordinates of the centre of the centre box in the window as a centroid of a cluster if it has the maximum number of data points compared with the rest.

After the above process, both the number of clusters and the estimated location of the centroids are obtained.

Pseudocode and Complexity.

```

# PREPROCESSING
let N = number of data points          # assignment, O(1)
let t = tolerance level, a fraction    # assignment, O(1)
let r = decrease factor, between 0 and 1
                                         # assignment, O(1)

Find max and min of the data set in both dimensions
      # finding the max two times and
      # min two times, 4*O(N)

Get bounding box by comparing max and min for both dimen-
sions
      # assigning max - min twice, 2*O(1)

let M = number of pixels of the longer side of the bound-
ing box
      # finding the larger number among
      # 2 numbers, O(2)

# FINDING OPTIMAL GRID SIZE
Let m = M                               # assignment, O(1)

DO:
/* DO-WHILE LOOP:
  number of iterations for m to reach a grid-size of 1
  (absolute worse case) given a predetermined decrease
  factor r, O(ln(M)/-ln(r))
  */
  m = m*r                                # assignment, O(1)
  Map number of data points into a m*m array by divid-
  ing the side of bounding box by m
      # going through each data point, O(N)
  Check number of grid boxes with only one data point
      # going through each grid box, O(M^2)
  WHILE (number of boxes containing only a single data
  point is under N*t)  # boolean checking, O(1)

# FINDING LOCAL DENSITY MAXIMA
For each grid box from the top-left to bottom-right by a
3*3 moving window:
      # going through each grid box, O(M^2)
  check maximum number of data points among the 9 grid
  boxes in the window  # going through 9 boxes, O(9)
  Check whether the centre box has the maximum number
  of data points      # boolean checking, O(1)
  Store the box if yes # assignment, O(1)
Count number of identified local maxima
      # going through all stored local
      # maxima, worse case O(M^2)
Return the number of local maxima and their coordinates
      # termination of algorithm

```

The number of iterations of the DO-WHILE loop can be derived by solving for x :

$$\begin{aligned}
 Mr^x &\leq 1 \\
 \ln(N) + x\ln(r) &\leq 0 \\
 x\ln(r) &\leq -\ln(M) \\
 x &\geq -\frac{\ln(M)}{\ln(r)}
 \end{aligned} \tag{1}$$

As $0 < r < 1$, $\ln(r)$ would be negative. So the inequality sign is flipped in the last line in Eq. (1).

For processes that involved the grid boxes, the worse case was assumed to be the smallest grid size of side 1 which gave M^2 number of steps.

The overall complexity will be:

$$\begin{aligned}
 &O\left(3 + 4N + 5 + \frac{\ln(M)}{-\ln(r)}[1 + N + M^2 + 1] + M^2(11) + M^2\right) \\
 = &O\left(\frac{\ln(M)}{-\ln(r)}[N + M^2 + 2] + 11M^2 + M^2 + 4N + 9\right) \\
 = &O\left(\frac{\ln(M)}{-\ln(r)}[N + M^2 + 2] + 12M^2 + 4N + 9\right)
 \end{aligned} \tag{2}$$

Since r is a predefined value that can be fixed across different datasets, it is considered to be a constant. $\ln(r)$ is negative and therefore $-\ln(r)$ can be dropped altogether as a positive constant. After dropping constants,

$$\begin{aligned}
 &= O(\ln(M)[N + M^2] + M^2 + 4N) \\
 &= O(N\ln(M) + M^2\ln(M) + M^2 + 4N) \\
 &= O(N[\ln(M) + 4] + M^2[\ln(M) + 1]) \\
 &= O(N\ln(M) + M^2\ln(M))
 \end{aligned} \tag{3}$$

3 Results

See Figs. 1, 2, 3, 4 and Table 1.

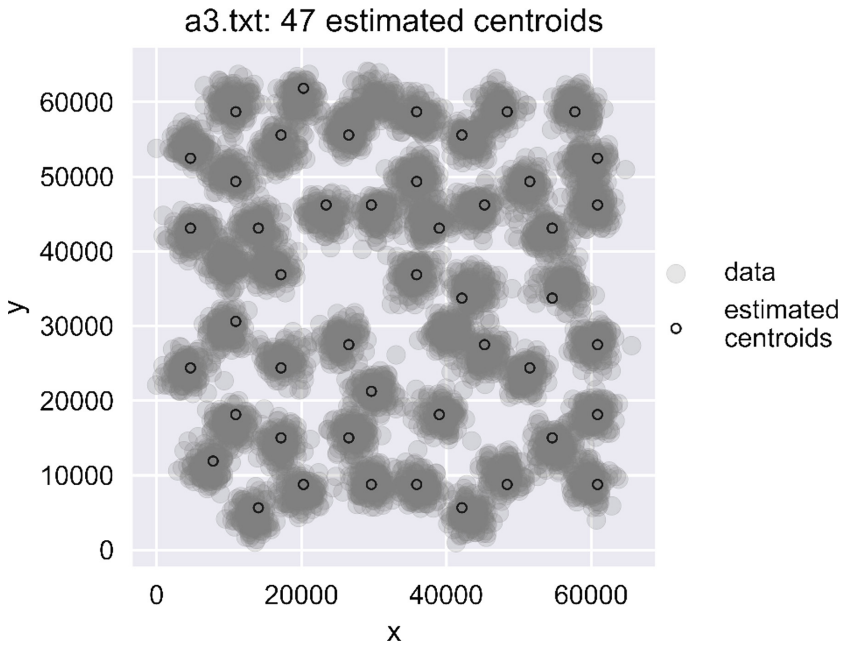


Fig. 1. Application of algorithm on a3 dataset

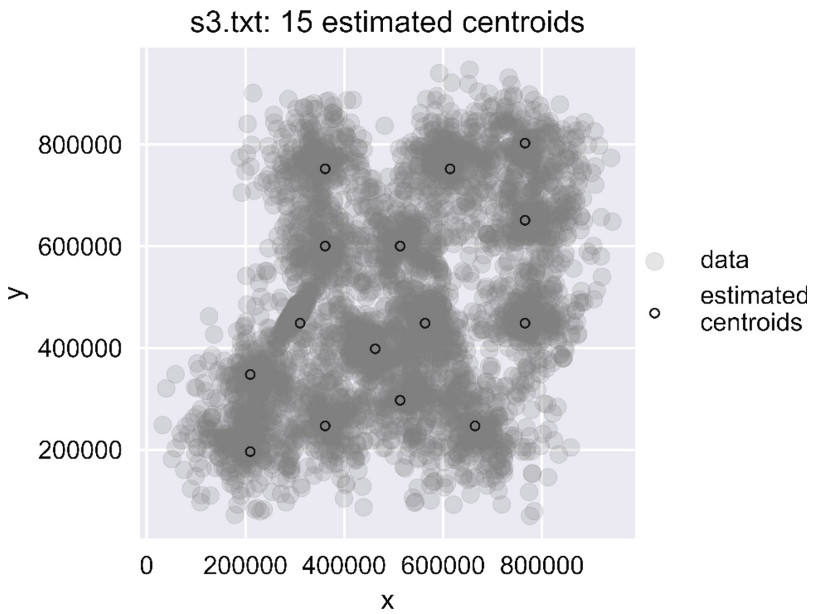


Fig. 2. Application of algorithm on s3 dataset

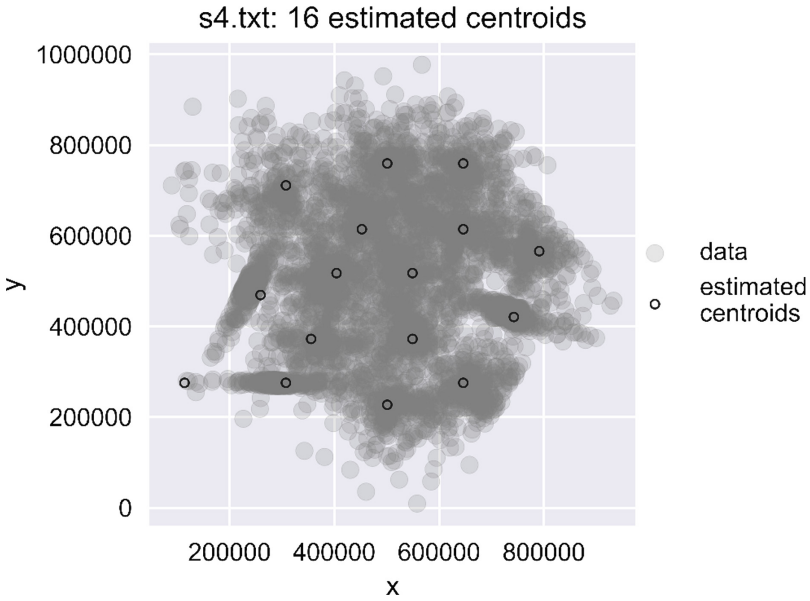


Fig. 3. Application of algorithm on s4 dataset

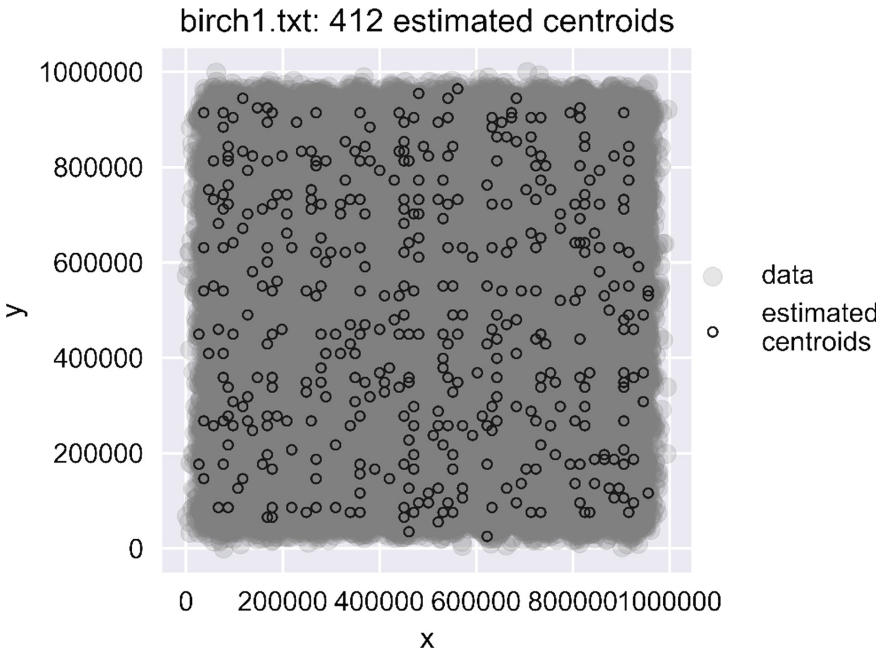


Fig. 4. Application of algorithm on Birch1 datasets

Table 1. Side-by-side comparison of ground truth k and estimated k

Dataset	Ground truth k	Estimated k of algorithm	Deviation from ground truth	Total run time (sec)	Time for finding optimal grid size (sec)	Time for centroid evaluation and value of k (sec)
a3	50	47	3	11.50498	11.47806	0.00409
s3	15	15	0	11.46858	11.44938	0.00263
s4	15	16	1	10.88031	10.86314	0.00296
<i>Birch1</i>	<i>100</i>	<i>412</i>	<i>312</i>	<i>222.207703</i>	<i>221.70673</i>	<i>0.07571</i>

4 Discussion

4.1 Assumptions

We make the following assumptions for the dataset. First, the dataset displays Gaussian distribution or near normal distribution. Second, clusters have a low degree of overlap. In other words, there is a certain separation measure between clusters. After all, human vision could also start to have difficulty identifying clusters with a high degree of overlap. To quantify this, in our study we used a standard deviation of no more than 42% distance between centroids, at a tolerance level of 0.5%.

4.2 Feature

As a pre-processing step that contributes to clustering, our algorithm offers the following advantages:

Robustness to Varying Densities. Our algorithm performed well on dataset S-3 where each cluster has varying densities, identifying all 15 ground truth clusters and finding centroids very close to the ground truth.

Robustness to Clusters of Different Shapes and Sizes. Our algorithm proved to be robust in cases where clusters are of different shapes. In the case of dataset S-4, one-third of the ground truth clusters are elongated as opposed to circular. Our algorithm identified all 15 ground truth clusters, although it found an extra cluster having been distracted by some outliers.

Guaranteed Termination of Algorithm in Polynomial Time. Our method of grouping the data into subregions and mapping the data in each region means that all data points are considered in the search for local maxima. The search will stop once the dataset is covered and the cut-off tolerance is reached. The guaranteed termination can help to limit the runtime and to ensure the algorithm reaches optimal accuracy.

Also, the complexity of the algorithm is $O(N \ln(M) + M^2 \ln(M))$, which is bounded above by $O(X^3)$, where X is the larger one between N and M . It is also worth noting that it is a very conservative estimation of the complexity as the worse cases for finding the optimal grid size and finding local maxima are considered separately without taking

into account the trade-off between the two processes. If a dataset requires a smaller grid size, it will need less steps in finding M but gives a larger value for M . This means the actual worst complexity of the algorithm will always be less than $O(N \ln(M) + M^2 \ln(M))$.

Other factors like the density of the data points, amount of noise and the shape of the clusters will also affect the actual run time, but their influences cannot be properly quantified as they will be different for individual dataset depending on the nature of each of them.

4.3 Critique

The main strength of our algorithm is perhaps its intuitive simplicity. On the flip side, complex cases and edge cases might not be the forte here.

Our initial assumptions of Gaussian distribution and non-overlapping clusters are also limitations. When the dataset presents some unusual situations violating our assumptions, our algorithm would underperform. For instance, where a single cluster contains more than two local maxima and displays varying densities at the same time, our algorithm cannot handle this complexity. If the cluster sizes vary radically, our algorithm is also not ideal.

When dealing with more complex cluster shapes, our grid can misalign with the cluster shape, resulting in less accurate centroid estimation. While this may be an acceptable margin of error for a pre-processing step, this is certainly an area for improvement.

4.4 Possible Improvements/Further Research

Clarity of Clusters. After testing our algorithm on the Clustering Datasets, several areas for improvement have been revealed. First of all, a score may be needed to indicate how clear-cut the clusters are. This is again applying a concept in human cognition to enhance the algorithm – a human can intuitively tell whether an image is clear or not. This would become more important when we try to apply the methods used here to higher dimension data.

Underclustering. A related issue is underclustering, which we saw in some cases such as seen in the Birch 1 dataset. This could be due to radical differences in cluster size, density or shapes within a given dataset. Even so, it would be beneficial to optimize the algorithm for dealing with outliers or noise.

Optimal Grid Size. Another area for further investigation is how to determine the optimal grid size. Our current method sets a tolerance level. It is worth exploring how we can use a method that is less arbitrary or less based on trial-and-error to determine the appropriate grid size.

Improving Computational Efficiency, Optimizing Exponential Step. The exponential step, currently set at 0.975, is thorough for our purposes but leads to numerous iterations. However, a larger exponential step may miss the optimal grid size.

We would need to find a way to decrease the number of iterations, especially at higher dimensions. This is very important as the runtime for finding the optimal grid size takes up over 95% of the total runtime for the current algorithm.

4.5 Value

As unsupervised learning is often highly dependent on input parameters from the user and the dataset's ground truth is not available, our algorithm is an attempt to apply human cognition to help. To a certain extent, our algorithm has shown to be informative, taking out the guesswork of the appropriate number of clusters to test for a clustering algorithm.

Given its robustness to differing cluster densities and shapes, although only in two dimensions, there may be potential in extending our method to higher dimensions. In addition, by indicating a suitable grid size, our algorithm can also be used to provide a good standard deviation value for kernel density clustering.

5 Conclusion

Grouping data into sub-regions provides valuable insights into the data distributions. By searching for local maxima in data densities, it can even give an estimation of number of clusters and centroid estimation to a satisfactory level of accuracy of maximum deviation of 3 data points if the data follows a Gaussian-like distribution and with SD lower than 42.4% of distances to the nearest neighboring cluster. Further research is needed to provide better results and more insights into human cognition in all levels will bring great value to the development of machine learning algorithms and artificial intelligence.

References

1. Estivill-Castro, V.: Why so many clustering algorithms: a position paper. *ACM SIGKDD Explor. Newsl.* **4-1**, 65–75 (2002)
2. Wong, K.-C.: A short survey on data clustering algorithms (2015). arxiv.org/pdf/1511.09123.pdf
3. Jain, A.K., et al.: Data clustering: a review. *ACM Comput. Surv. (CSUR)* **31**(3), 264–323 (1999)
4. Yadav, J., Sharma, M.: A review of K-mean algorithm. *Int. J. Eng. Trends Technol. (IJETT)* **4**(7), 2972–2976 (2013)
5. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: part I and part II. *ACM SIGMOD*, **31**(2), 40–45 and **31**(3), 19–27 (2002a, 2002b)
6. Kriegel, H.-P., Kröger, P., Sander, J., Zimek, A.: Density-based clustering. *WIREs Data Min. Knowl. Discov.* **1**(3), 231–240 (2011)
7. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008)
8. Kohonen, T., Honkela, T.: Kohonen network. *Scholarpedia* **2**(1), 1568 (2007)

9. Bação, F., Lobo, V., Painho, M.: Self-organizing maps as substitutes for k-means clustering. In: Sunderam, V.S., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2005. LNCS, vol. 3516, pp. 476–483. Springer, Heidelberg (2005). https://doi.org/10.1007/11428862_65
10. Wong, K.-C., Peng, C., Li, Y., Chan, T.-M.: Herd clustering: a synergistic data clustering approach using collective intelligence. *Appl. Soft Comput.* **23**, 61–75 (2014)
11. Tan, F., Slotin, J.-J.: A quorum sensing inspired algorithm for dynamic clustering (2015). arxiv.org/pdf/1303.3934.pdf
12. Santosh, D.: Tracking multiple moving objects using gaussian mixture model. *Int. J. Soft Comput. Eng.* **3**(2), 114–119 (2013)
13. Han, S., Humphreys, G.W., Chen, L.: Uniform connectedness and classical gestalt principles of perceptual grouping. *Percept. Psychophys.* **61**, 661–674 (1999)



Human-Machine Teaming and Cyberspace

Fernando J. Maymí¹ and Robert Thomson²(✉)

¹ Soar Technology, Ann Arbor, MI 48105, USA
fernando.maymi@soartech.com

² Army Cyber Institute, West Point, NY 10996, USA
robert.thomson@usma.edu

Abstract. Artificial Intelligence is becoming the key enabler of solutions to a variety of problems including those associated with cyberspace operations. Based on our analysis of cyber threats and opportunities in the coming years, we assess it as very likely that teams consisting of humans and synthetic agents will routinely work together in many if not most organizations. To fully leverage the potential of these teams, we must continue to develop new paradigms in human-machine teaming. Specifically, we must address three areas that are currently in their infancy. Firstly, we need interfaces that allow all teammates to communicate effectively with each other and seamlessly transfer tasks among them. This must be true regardless of whether the endpoints are human or not. Secondly, we will need cybersecurity operators with broad knowledge and skills. They must know how their synthetic teammates “think,” when to task them and when to question their reports. Thirdly, our AI systems must be able to explain their decision-making processes to their human teammates. This paper provides an overview of cyberspace threats and opportunities in the next ten years and how these will impact human-machine teaming. We then apply the key lessons we have learned while working a multitude of advanced research projects at the intersection of human and AI agents to cyberspace operations. Finally, we propose areas of research that will allow humans and machines to better collaborate in the future.

Keywords: Human-machine teaming · Artificial intelligence · Cyberspace

1 Introduction

The United States Department of Defense (DoD) defines cyberspace as a global domain consisting many different and often overlapping networks (Joint Pub 3-12 2013). Though many people equate cyberspace with the Internet, the latter is simply a subset of the former. Cyberspace, after all, includes many networks (e.g., classified intelligence networks) and systems that are not directly reachable from the Internet. Though it is difficult to characterize the nature of these other networks and systems that comprise cyberspace, we know a fair amount about the Internet. We know, for instance that it is the largest, most complex system ever built by humans. By some estimates, it consists of over 8 billion devices (Tung 2017) exchanging over 4 billion bytes of data every second (“Internet Live Stats” 2018). Cyberspace, by definition, is even bigger.

This is a U.S. government work and its text is not subject to copyright protection in the United States; however, its text may be subject to foreign copyright protection 2018

D. D. Schmorow and C. M. Fidopiastis (Eds.): AC 2018, LNAI 10915, pp. 299–315, 2018.
https://doi.org/10.1007/978-3-319-91470-1_25

The size and speed of the Internet, coupled with its growth rate, prompted the development and application of artificial intelligence (AI) techniques for performing tasks that humans alone could no longer effectively do at this scope. It also spurred the creation of novel capabilities that take advantage of, and indeed require, the very large data sets that are available in cyberspace. The development of these techniques has been organic and, while enabling localized capabilities, has sometimes hindered other ones. In particular, we are concerned that some trends in both human and synthetic (i.e., AI-enabled) operator development are not supportive of effective human-machine teaming. In Sect. 2 of this paper, we provide a brief introduction to AI in general and to some of the specific concepts we'll discuss later in the paper. On this foundation, we describe in Sect. 3 future threats that motivate the need for better human-machine teaming. We then describe advances in AI that allow the creation of synthetic cyberspace actors in Sect. 4. In Sect. 5, we address the human members of future cyberspace operations teams. Section 6 presents the need for AI that is explainable to humans as the foundation of trust in these teams. These human-machine teams of the future are described in Sect. 7. Finally, we offer our conclusions and recommended future work in Sect. 8.

2 A Brief Introduction to AI

AI is fundamentally concerned with machines that solve problems and make decisions or appear to think analogously to a human at some level of approximation. While there is no single definition for the term, there exist different classes of AI that allows us to formulate a tentative ontology, which we show in Fig. 1. A high-level bifurcation is possible by differentiating the approach used to represent information or knowledge. Symbolic approaches, as the name implies, use symbols (e.g., words) to represent the atomic components of thought and generally rely on some kind of semantic rules to process information. Alternately, non-symbolic approaches use numerical and often distributed representations (reflected by patterns of activity across numerous processing units).

In symbolic approaches to AI, system developers model real-world concepts, their relationships and how they interact to solve a set of problems. This effort requires considerable knowledge of both the problem and solution domains, which makes it fairly labor-intensive. However, it yields results that are inherently explainable to humans since they are derived from human knowledge models in the first place. Symbolic AI systems include the expert systems that became prolific in the 1970s and 80s. These relied on extensive interviewing of subject matter experts and time-consuming encoding of their expertise in a series of conditional structures. An example of this approach is MYCIN, one of the first practical rule-based systems that was developed to help physicians select antimicrobial therapies (Shortliffe 2012). These early expert systems suffered from a fundamental inability to adapt or learn absent human intervention in updating the knowledge base.

Non-symbolic AI gained momentum after many in the AI community, disappointed with the limitations of symbolic approaches, looked to animal brains for inspiration. In Artificial Neural Networks (ANN) each node receives multiple inputs from other

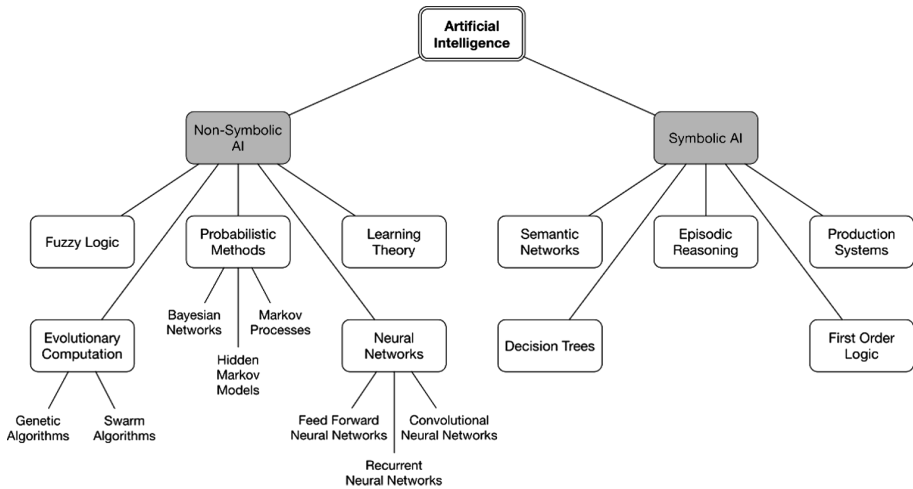


Fig. 1. A general ontology of AI techniques

nodes, typically in the form of a real number, and produces one or more outputs that are the result of applying some function to those inputs. By applying weights to each connection and allowing those weights to be modified through a feedback loop, the ANN can be trained. There are many other non-symbolic approaches, such as probabilistic ones, that have been successfully applied to problem sets in which the knowledge engineering required in symbolic AI is not a feasible option.

Many modern AI systems are able to learn from experiences. Machine learning (ML) refers to techniques that allow AI systems to adapt to changing inputs and ideally improve their performance as a result. Though ML is equated with non-symbolic approaches, it is also possible for some symbolic AI systems to learn. The Soar cognitive architecture, for instance, is a symbolic production system capable of episodic learning. A Soar agent could achieve a goal through a circuitous series of intermediate steps, some of which will be successful. Over multiple experiences or episodes, the agent condenses these steps into a shorter, more effective and efficient chain. This process, called “chunking,” is one of the main ways in which Soar agents learn.

ML can take place with or without human help. In supervised ML, the system is presented with inputs and must then produce an output, typically in terms of a classification (e.g., an email message is or is not spam). If the output is correctly classified, the system receives positive reinforcement; otherwise, it may receive negative reinforcement. The learning or, more accurately, training process can be automated by using labeled training data sets. If you remove the human from the process and don't use labeled data, a system can still learn through reinforcement learning. In this form of ML, the system interacts with its environment in a sequence of observation-action pairs where a reward is presented after each action. Using this approach, a system could learn how to efficiently route network packets using as rewards the inverse of the number of hops required. The key requirement for these ML approaches is a feedback process that allows the agent to determine when its decisions are correct. This process can be

artificial (e.g., tagged data sets) or natural (e.g., observing the behavior of routed packets).

3 Future Threats

The ability of AI to make increasingly complex decisions much faster than humans could, all the while learning from its experiences has already delivered many benefits in the service of humanity. The same capabilities, however, can cause unexpected and undesirable effects as Microsoft learned when it developed chatbot that learned to compose racially and sexually offensive tweets (Metz 2016) from its interactions with thousands of people. Perhaps more concerning are scenarios in which AI systems are intentionally developed and deployed to cause harm. It is, after all, logical to assume that malicious cyberspace actors will leverage emerging technologies for their own purposes.

If an attacker is using AI to operate at machine speed, the defender must be able to work at least as quickly in order to be effective. This idea of synthetic agents attacking and defending information systems with no humans in the decision-making loop inspired the Defense Advanced Research Projects Agency (DARPA) Cyber Grand Challenge (CGC), which brought together seven finalists to Las Vegas, Nevada in August of 2016. The goal was for these cyber reasoning systems (CRS) to perform automated vulnerability detection, exploit generation, software patching, and to determine when it would be most advantageous to patch a vulnerability or exploit it on a competing team's CRS without human intervention (Brooks 2017). The message is clear: in the future of cyberspace both attackers and defenders will, at least partially, be autonomous agents. In fact, the leader of the winning team, David Brumley, founded the company For All Secure to take autonomous vulnerability detection (and potentially patching) to market.

It is not only machines who will be threatened by autonomous agents. Many security experts anticipate a new breed of phishing emails generated by ML algorithms that will be much more targeted, compelling and effective than human-generated ones (Emmanuel 2017). One of the reasons why these messages will be more threatening is that they will leverage the ability of data analytics and ML to scour vast data sources for information with which to precisely target individuals at scale. The U.S. Army already identified this micro-targeting trend as a feature of future wars (Kott et al. 2015) for which our current counter-measures are ineffective.

The Social Network Automated Phishing with Reconnaissance (SNAP_R) system (Seymour and Tully 2016) demonstrated a recurrent neural network (RNN) that is able to tweet phishing messages that target specific users. During a limited experiment, SNAP_R was four times faster than humans at sending out targeted attacks while achieving an order of magnitude improvement in the target click rate. One year later, DARPA announced its Active Social Engineering Defense (ASED) program, aimed at autonomously identifying, disrupting, and investigating social engineering attacks. The very existence of ASED underscores the difficulty and long-term significance that DARPA attributes to this threat.

Finally, as AI in general and ML in particular become increasingly important in our lives, adversaries will develop attacks aimed directly at the ML mechanisms that are designed to improve and defend our lives. Adversarial ML (AML) is an emerging field of study concerned with attacks against online ML algorithms (Huang et al. 2011). Early research has shown that ML classifiers are susceptible to three types of attacks (Papernot et al. 2016). Confidentiality attacks entail gaining information on data used to train the ML system (Shokri et al. 2017), its internal model (i.e. weights), or architectural (i.e. learning rate) parameters. Integrity attacks attempt to modify input to the ML classifier in order to induce a particular output or behavior, such as causing an image recognition system to misclassify a 2 as a 9 by modifying a few image pixels (Carlini and Wagner 2017). Availability attacks attempt to deny access to the ML classifier such as by generating numerous false positives. An ML-based IDS/IPS, for example, is vulnerable to such attacks. As AML techniques mature, malicious actors will employ them to manipulate the outputs of intelligent systems.

4 Synthetic Actors

Against this backdrop of technological opportunities and threats, research and development of autonomous synthetic actors proceeds apace. Though much work to date has focused on applications of ML to the detection and mitigation of cybersecurity incidents, research is also taking place towards the development of more robust defensive agents that can hunt for and neutralize threats on their networks. As we mentioned in our threat discussion, we are seeing similar moves in the development of attack capabilities. In fact, one of the noteworthy aspects of DARPA's CGC was that it demonstrated the feasibility (and, one might argue, inevitability) of autonomous offensive and defensive agents fighting against each other with humans out of the loop at least in some cases. While we have not yet seen documented cases of autonomous synthetic attackers conducting real operations, many think that these incidents are not too far in the future (Dvorsky 2017). Indeed, SoarTech has already demonstrated cognitive agents that can perform defensive and offensive (e.g., penetration testing) activities in virtualized environments.

SoarTech's Simulated Cognitive Cyber Red-team Attacker Model (SC2RAM) is a synthetic, offensive, cognitive agent that emulates real attackers by modeling the complex thoughts, decision-making, and contextual understanding of a human interactive operator. Its goal-seeking behavior results in a virtually unlimited range of realistic attacks. The current attacker agent, built on the Soar Cognitive Architecture (a symbolic AI platform) can conduct multiple attacks including phishing with malicious documents, remote exploitation, and SQL injection. A custom remote access toolkit developed for this project provides additional persistent on-target capabilities such as lateral movement and file exfiltration, providing a realistic experience for training network defenders. The premise of red teaming and penetration testing, exemplified by SC2RAM, is that it is better to test one's own defenses against realistic but benign attackers than it is to wait until the real adversaries do so. Since human penetration testers are rare and expensive experts, it is logical to leverage synthetic agents in this manner.

It also makes sense to employ such agents when the scale of a problem requires a very large number of interactions. Much of the research at the intersection of cybersecurity and AI uses non-symbolic approaches. Some of the first successful applications of ML to cybersecurity were in classification of spam email messages (Cohen 1996). Over the last two decades, these approaches have become remarkably accurate. Today's ubiquitous spam filters improve their performance through interaction with the humans whose inboxes they protect. When the agent misclassifies a message, the human has an opportunity to correct the error thus allowing the ML system to learn to improve itself.

Given the role of these agents as first lines of defense for end points, much research is needed in identifying vulnerabilities to AML in systems such as these spam filters or the newer breeds of antimalware products that use ML to detect malicious software. Here one could utilize machine learning techniques to make inferences on the training set of another machine learning classifier in order to manipulate inputs to generate desired outputs. For example, given an ML system that classifies software as benign or malicious (e.g., an anti-malware application), one could imagine another system that generates multiple variants of malware, each with small perturbations that don't affect its functionality. These variants could be sent to the classifier until it incorrectly decides that the malware sample is benign. Given enough such misclassified samples, the AML system can make inferences about what it takes to fool the defender. This AML versus ML assessment could serve to harden network security applications by evaluating the robustness of an already trained model, particularly when the internal classifier parameters are unknown. Since this sort of assessments require many thousands or millions of attempts to characterize the system under test, synthetic agents would be well-suited to perform them.

Despite their ability to analyze vast amounts of information, non-symbolic approaches like those in use for spam, malware and intrusion detection are less effective at reasoning over the context and meaning of cyberspace activities. They are ideally suited to answer the questions of *what* and even the *how*, but not the *why*. Symbolic approaches, such as rule-based systems, on the other hand, are oftentimes better for this purpose because they model higher-level cognitive processes and human expertise. A promising area of research for more effective synthetic cyberspace actors is the integration of symbolic and non-symbolic approaches to help us identify not just the threats, but also their possible implications to our organizations and systems. Such hybrid systems would be more capable in a wider variety of situations. It will be at that point that synthetic actors could become real teammates to their human counterparts, significantly enhancing the performance of our workforce.

5 Human Actors

One of the challenges in reviewing the current state of the cyber workforce is that there is a paucity of quantitative assessment regarding the cognitive aptitudes, work roles, or team organization required by cyber professionals to be successful. We argue that the people who operate within the cyber domain need a combination of technical skills,

domain specific knowledge, and social intelligence to be successful. They, like the networks they operate, must also be secure, trustworthy, and resilient.

A concern in writing about human actors is that cyber professionals are generally seen as a homogeneous, holistic classification. That said, due to the complexity and rapid evolution of the tasks involved in cyber defense, it is important to note that there is substantial heterogeneity between work roles and individual skillsets. By virtue of this complexity in the task environment, cyber professionals need to work in teams. While in the military context cyber teams tend to be teams of diverse talents, in the private sector it is much more likely for smaller teams to be composed of similarly-talented individuals rather than a group with diverse work roles and backgrounds (Champion et al. 2012). Recent research has identified that cybersecurity teams are better able to solve complex tasks than individual analysts, potentially due to the distribution of expertise across analysts (Rajivan 2014; Rajivan et al. 2013; Rajivan and Cooke, in press). For instance, performance on incident triage was highest with a diverse group of heterogeneous talents as opposed to a team with members of similar background and skills. (Rajivan 2013). A limitation of research into cyber teamwork is that they have not examined different organizations of teams or combinations of teams. This future research is essential to determine the correct make-up of the future cyber workforce.

Champion et al. (2014) investigated the contribution of informal education to developing cyber security expertise and found that 69 of 82 professionals reported that informal education supplementation was a prerequisite for career success. Furthermore, 40% of professionals felt that job experience was the highest factor in positive performance over degree of knowledge/education (12%). Many professionals anecdotally reported that those receiving supplemental on-the-job training and mentoring exhibited the highest performance benefits as measured by future career success. Similarly, Asgharpour et al. (2007) found that operators who subjectively rated themselves with higher levels of expertise tended to have both more and more diverse competencies than those with less self-professed expertise.

Cognitive task analyses have identified that cyber professionals need to exhibit strong situational awareness (Jajodia et al. 2010), including juggling concurrent sources of information regarding the health of the network, historical and current network activity, and performing a continual assessment of risk. For recent meta-analyses see Franke and Brynielsson (2014), and Onwubiko and Owens (2011). Similarly, through the use of structured interviews, Goodall et al. (2009) interviewed twelve cyber professionals and identified that the requirement for situated knowledge (i.e., knowledge of the local environment) made intrusion detection a relatively unique task and challenging to transfer expertise to other tasks in the cyber domain. This required triage teams to interface with local workers to understand the topology and peculiarities of the local network to determine whether an intrusion had occurred and what remedies were available.

There are numerous tools to process this incoming information (e.g., Bro and Snort for intrusion detection), however, there is just too much information for a human actor to successfully process, and critical misses are inevitable. A human teamed with a machine, however, has the potential to cover a much wider set of attack vectors

because the machine does not have the same attentional limitations and can do a more thorough assessment of making sense of large swaths of incoming data.

Before proceeding to discuss the importance of AI systems that can interact with human actors, it is important to understand how we are training our cyber workforce and to identify any gaps in training. The Department of Homeland Security's National Initiative for Cybersecurity Careers and Studies (NICCS) developed a Cybersecurity Workforce Framework (Newhouse et al. 2016) to provide a base set of work roles for the cyber workforce. While this ontology was not empirically justified, it represents the most well-documented rostering of work roles in the cyber domain. This collection includes nine work-role categories, 31 specialty areas, and over 1000 types of knowledge, skills, and abilities. Major categories are described in Table 1.

Table 1. Cybersecurity Workforce Framework. Reproduced from (Newhouse et al. 2016, p. 14).

Work-role category	Description
Securely provision	Conceptualizes, designs, and builds secure information technology (IT) systems, with responsibility for aspects of systems and/or networks development
Operate and maintain	Provides the support, administration, and maintenance necessary to ensure effective and efficient information technology (IT) system performance and security
Oversee and govern	Provides leadership, management, direction, or development and advocacy so the organization may effectively conduct cybersecurity work
Protect and defend	Identifies, analyzes, and mitigates threats to internal information technology (IT) systems and/or networks
Analyze	Performs highly specialized review and evaluation of incoming cybersecurity information to determine its usefulness for intelligence
Collect and operate	Provides specialized denial and deception operations and collection of cybersecurity information that may be used to develop intelligence
Investigate	Investigates cybersecurity events or crimes related to information technology (IT) systems, networks, and digital evidence

Securely Provision roles revolve around the more traditional information technology field including software developers, computer programmers, and network architects. The Operate and Maintain roles include System Administrators, Knowledge Management, and Security Analysts. The Oversee and Govern roles include managerial roles, Cyber Law, Policy Development, and Education. The Protect and Defend roles include Cyber Analysts (Operators) and Network Defenders. The Analyze, Collect and Operate, and Investigate roles all encompass the broad field of Digital Forensics and will tend to be government or law enforcement positions (Caulkins et al. 2016).

In general, cyber professionals in the Securely Provision, Operate and Maintain, and Protect and Defend work roles must have good mental flexibility and pattern matching abilities (Baker 2016; Ben-Asher and Gonzalez 2015; Champion et al. 2014).

They will have to possess significant skill and knowledge about computer operating systems and using analytical tools for such things as network scanning, network mapping, and vulnerability analysis. This task environment involves scanning large numbers of network events and (generally false) alerts across multiple computer screens with the goal of identifying threats while minimizing false alerts (D'Amico and Whitley 2008).

A limitation of the NICCS Workforce Framework is that, of the 1060 types of knowledge, skills, and aptitudes, fewer than ten describe teamwork or working with AI. This implies that the Framework paints an incomplete picture of workforce proficiency (Cook 2014). Furthermore, the development of any cyber workforce that neglects the social aspect of human behavior on the network neglects a critical component of the cyber domain. For instance, cyber defense would be aided by an understanding of human behavior and how it introduces risk to the network (Asgharpour et al. 2007; Pfleeger and Caputo 2012). We should leverage AI and humans' capabilities to maximize information exchange so each level processes the right 'kinds' of information to be most effective. Under this view AIs should process the large swaths of incoming poorly-structured data and distill this data into a format that can be readily presented to a human operator. The human operator can then perform high-level strategic inference over this well-structured information from the AI. We now know that human operators, though, will not use this data unless they can understand why the AI makes its recommendations.

6 Explainable AI

Most AI systems today are not designed to (nor can they usually) explain to their human users the manner in which they arrived at their conclusions. The reason is that most AI developed to date for cybersecurity applications is non-symbolic. As we explained in our introduction to AI earlier, these approaches, unlike symbolic ones, are not inherently explainable. System designers would have to deliberately develop explanation mechanisms, which is something seldom seen in the field. Faced with such opacity, many users choose to blindly trust the computer, which is a phenomenon that has been called the "in screen we trust" effect (Aiken 2017). The option is to distrust the computer and ignore its decisions if they seem unreasonable. Some systems, however, might not allow this option if their AI mechanisms are part of closed decision loops that don't allow real-time human interference.

In order to develop and maintain the trust that is inherent in teaming, AI systems must be able to explain their conclusions to human teammates. In this regard, symbolic AI approaches such as expert systems and cognitive architectures are better suited because they model human knowledge and thought processes respectively. Their very nature is similar to higher level human thought constructs, which in most cases makes it simpler for them to present their causal chains to humans. Conversely, this nature also makes it easier for humans to point out errors or omissions in their synthetic teammates. The Soar cognitive architecture, for instance, uses goal graphs to simulate human cognitive processes, which naturally lends itself being explainable to people by representing the synthetic decision-making processes as goal trees.

Visualizing non-symbolic systems like ML processes, on the other hand, has traditionally been more difficult. The reason for this is that they mostly rely on mathematical models and processes. To this end, the Defense Advanced Research Projects Agency (DARPA) is pursuing its eXplainable AI (XAI) program, to which the authors of this paper are both contributing. One of projects in this program is XAI for the Veterans’ Transition Assistance Program (XAI-VTAP), which is geared towards matching the resumes of veterans to open job postings. Some of the work being done in this project uses novel techniques to provide an unprecedented level of visibility into how ML algorithms arrive at conclusions. Figure 2 shows how and why the system matched a specific resume to multiple occupational categories. The top part of the figure shows how good of a fit a candidate is against each category and provides examples from that person’s resume. The bottom part illustrates how various indicators were ultimately mapped to various categories. One could imagine job seekers using this feedback as a training aid to create better resumes in general, as well as resumes that improve their odds of getting specific jobs.

Occupational Category	Maximum Activation	Occurrences in Resume	Example from Resume
general and operations managers	32.3	29	"employment history ai cyber rd director 2016 present us"
software developers, applications	60.8	13	"arduino pic developer tools eclipse vi ms visual studio svn git uml avr studio matlab octave artificial intelligence systems soar clips jess machine learning robotics player stage ros gazebo"
chief executives	18.3	9	"accomplished public speaker speaking in both technology focused and senior executive level sessions"
software developers, systems software	12.1	9	"automated army mobilization and deployment processes by architecting and overseeing development of several multi million dollar distributed cloud based applications"
information security analysts	16.2	7	"technical experience in cyber security autonomous systems cognitive systems and computer science and engineering research development and education"

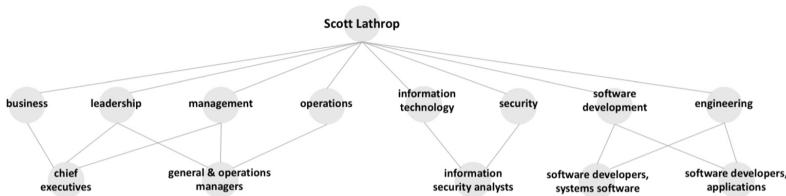


Fig. 2. User interface prototype for Explainable AI to support Veterans Transition Programs (XAI-VTAP)

While such explainability could lead to better employment opportunities and possibly improve resume-writing skills for veterans and other job seekers, it also enables the threats to AI systems posed by AML. While there are many types of AML, the one that is most relevant to our discussion is the deliberate manipulation of the data inputs to an ML mechanism so that it fails to function as intended. This could happen if an adversary determines how an ML-based spam filter works and then crafts spam messages that are not identified as such and thus are delivered to a victim’s inbox. It could also happen if the adversary pollutes the training data set for an ML-based product so that it is trained to correctly identify spam messages except those that have a particular set of characteristics that only the adversary knows. This would then allow

only that adversary to bypass the detection mechanism. The knowledge that can be gained through explainable AI facilitates AML techniques.

Still, explainability is crucial to our human-machine teaming efforts for three reasons. Firstly, it allows trust to be developed between humans and their synthetic teammates. Autonomous AI agents are likely to reach some seemingly far-fetched conclusions that may stretch their credulity of their human counterparts. In those situations, it is necessary to be able to walk the human through the thought process. Secondly, the AI system's conclusions will only be as good as the models they have and the learning they have been able to do on their own. It is entirely possible that some misfits may occur, in which case the human will be able to detect the error, point it out to the AI system, and allow it to learn from the experience. Thirdly, synthetic teammates have tremendous potential as training tools, which can only be realized if it is able to explain itself to those who are learning from or with that system.

7 Human-Machine Teaming

The notion of shared mental models between humans and machines is a common thread when examining human-centered big data research. Mental models provide a representation of situation, various entities, capabilities, and past decisions/actions. These models are dynamic, with analyst and model engaged in a continuous production loop. In addition, from a purely human level there is research on teamwork (Baumann and Bonner 2013) and the degree to which teammates from different backgrounds have overlapping shared mental models (Bearman et al. 2010). There is also research on the degree to which multiple agents can recognize a common plan from reading large corpora (Paletz 2014).

Teams of security analysts are in many instances, a loose association of individuals, rather than a functioning team (Champion et al. 2012). A functioning professional team is a “purposive social system” (Hackman and Katz 2010), in which members of the team have diverse backgrounds, identified by role and work together in an interdependent manner towards common objectives (Salas et al. 1992). Team effectiveness largely depends upon appropriate leadership, team structure, communication, collaboration and distribution of tasks. Communication is the key medium by which human teams form relationships, collaborate and share information (Cooke et al. 2013). Communication is the conduit to transform individual expertise and situational awareness to team level knowledge and situational awareness.

Field studies with security analysts found that communication and collaboration between security analysts was an integral aspect of effective defense particularly during a widespread security crisis (Goodall et al. 2009; Jariwala et al. 2012). Lab experiments on collaboration during the threat detection have also found evidence that cooperation between security analysts during triage analysis augments signal detection performance, particularly in novel and complex situations (Rajivan et al. 2013). However, during collaborative analyses, analysts may fail to contribute requisite expert knowledge and demonstrate biases in the way information is pooled from each other, leading to communication losses affecting threat detection performance (Rajivan 2014). Communication across the hierarchy of security analysts have also been observed to be

inefficient and largely one-directional (bottom-up). Tools for collaborative threat detection developed using human systems engineering principles would help in mitigating such losses in communication between security analysts (Rajivan 2014).

Leadership is also crucial to security defense team development and performance (Buchler et al. 2017). Typically, an individual in a leadership role is expected to: develop team capabilities, facilitate problem solving, provide performance expectations, synchronize and integrate team member contributions, clarify team member roles and engage in meetings and feedback (Salas et al. 1992; Simsarian 2002). Field studies on security leadership showed that leadership is a significant predictor of defense performance. In one such study, two security teams, otherwise equivalent in skills, experience and knowledge, was observed to demonstrate widely different defense performance primarily due to differences in leadership approach and amount of collaboration (Jariwala et al. 2012). In a subsequent study, it was found that functional specialization and adaptive leadership strategies are important predictors of security defense performance (Buchler et al. 2017). Except for these handful of studies, the determinants of effective teamwork and leadership among security analysts is still an emerging area.

Collaboration, communication and knowledge integration is necessary for accurate and expeditious correlation analysis. From past team research, it is evident that teams often don't realize their full potential and could fail for a multitude of reasons. Loss in team processes such as communication would lead to sub-optimal decision making. For example, collaborative threat detection requires the exchange of expert information between security analysts. Previous research has demonstrated that teams may not be effective in exchanging novel information. Particularly, uneven information distribution biases people to share, more often, information that are known to majority in the team and prevents them from sharing and associating unique information available with them (Stasser and Titus 1985). The effect of such team-level biases on security team collaborations are largely unknown.

Experiments on team interactions need to be conducted ideally in context (through field studies) or using simulation environments. Due to restricted access to real world cyber protection teams and due to lack of importance currently given to team process metrics in cyber defense exercises (Granåsen and Andersson 2016), experiments on team interactions in cyber defense can instead be conducted in the lab using simulation systems that recreate realistic team interactions and work flows between study participants which would in turn require the participants to exercise some of the same cognitive process involved while conducting cyber defense in the real world (Cooke and Shope 2004).

We argue that in order to incorporate machines into human teams effectively, they must be natural to use, seamlessly integrate into the task environment, and provide a subjective improvement in effectiveness. Ideally, a single human operator (or small team of operators) would be able to supervise multiple AIs (Chen and Barnes 2014; Pellerin 2015; Trexler 2017). The goal of the AI is to process the massive amount of incoming information, present it efficiently to the human operator, make low-level decisions, and help the human operator make high-level strategic decisions. This AI will be able to make decisions at the speed of cyberspace and adapt to new attack vectors in near real-time, which is orders of magnitude faster than a human operator.

We foresee that within the next decade, the war for cyberspace will be fought between nations' AIs, and the skill of the operators and effectiveness of the AI's algorithms will be the deciding factor.

As such, it is essential for human operators to trust their AIs. Petraki et al. (2015) argue that it is important to have mutual predictability and adaptability in order engender trust. As previously discussed, that is one of the main goals of DARPA's eXplainable AI Program. The ability of the AI to be able to adapt to a human operator's goals, and for the operator to query the underlying question as to 'why' a decision was made is key to trusting in the AI's automation. One such technique is to supplement traditional AI techniques with models that approximate human behavior, such as in the Soar cognitive architecture and the ACT-R cognitive architecture.

In summary, by leveraging AIs to do much of the complex sensemaking required in many cyber operations tasks, we argue that it is possible to maximize a human operator's ability to conduct strategic operations effectively, even in the face of an overwhelming amount of incoming data. We argue that AIs need to seamlessly integrate with humans, and that they need to be explainable in order for human teammates to trust their output.

8 Conclusions

From the foregoing, we posit that there are three key elements of effective human-machine teaming in cyberspace: effective intra-team communications mechanisms, a sophisticated and diverse cyber workforce, and AI systems that can readily explain the rationales for their decisions to their human teammates.

We have already established that communication is the key medium by which teams form relationships, collaborate and share information. It is a logical extension of this premise to assert that whatever the team composition (e.g., human, synthetic), as long as there is at least one human in the mix, effective communications will be required to build and maintain the team's effectiveness. Even if there are no humans in a team of cyberspace actors, communications will be key, albeit in a somewhat different form.

It will also be important to ensure that the human actors that are teaming with AI systems are knowledgeable of the capabilities and limitations of the underlying technologies. In other words, to fully leverage the potential of our synthetic teammates, we will need cybersecurity operators with broad knowledge and skills, and who know when to task agents and when to question their reports. There is a dearth of research in this area, so much work needs to be completed before we can quantify the requirements for humans in an effective human-machine cybersecurity team.

Finally, the skills of the human actors will be excessively tasked unless their synthetic teammates are able to explain to them the manner in which they reached a specific decision. This requirement for explainable AI addresses two critical aspects of effective teaming: trust and correctness. An important element of teamwork is trust, which can be eroded by unexpected behaviors, particularly those that could seem to undermine or threaten mission accomplishment. If a synthetic agent is incapable of explaining to its teammates how it arrived at a particular conclusion, it will not

engender (and may erode) trust. Furthermore, since it may likely be infeasible to develop a perfectly correct AI system, the ability to explain itself will allow its human teammate to identify logical or syntactical errors.

Given that it is likely that AI will play an increasingly important role in the future of cybersecurity, it is imperative that we develop better constructs for human-machine teaming. These should be focused on effective communications, human workforce development, and explainable AI. Though much research is needed in all three areas, we can't afford to take the risk of not getting this right. Our cybersecurity depends on it.

References

- Abbas, H., Petraki, E., Kasmarik, K., Harvey, J.: Trusted autonomy and cognitive cyber symbiosis: open challenges. *Cogn. Comput.* **8**(3), 1–24 (2015)
- Aiken, M.: *The Cyber Effect: A Pioneering Cyberpsychologist Explains How Human Behavior Changes Online*. Spiegel & Grau, New York (2017)
- Asgarpour, F., Liu, D., Camp, L.J.: Mental models of computer security risks. In: Dietrich, S., Dhamija, R. (eds.) *International Conference on Financial Cryptography and Data Security*. Lecture Notes in Computer Science, vol. 47886, pp. 367–377. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-77366-5_34
- Baumann, M.R., Bonner, B.L.: Member awareness of expertise, information sharing, information weighting, and group decision making. *Small Group Res.* **44**, 532–562 (2013)
- Baker, M.: *Striving for Effective Cyber Workforce Development*. Software Engineering Institute, Carnegie Mellon University, Pittsburgh (2016)
- Ben-Asher, N., Gonzalez, C.: Effects of cyber security knowledge on attack detection. *Comput. Hum. Behav.* **48**, 51–61 (2015)
- Bearman, C.R., Paletz, S.B.F., Orasanu, J., Thomas, M.J.W.: The breakdown of coordinated decision making in distributed systems. *Hum. Factors* **52**, 173–188 (2010)
- Brooks, T.N.: *Survey of Automated Vulnerability Detection and Exploit Generation Techniques in Cyber Reasoning Systems* (2017). arXiv preprint [arXiv:1702.06162](https://arxiv.org/abs/1702.06162)
- Buchanan, A., Goodall, L., Walczak, D'Amico, P.: *Mission impact of cyber events: Scenarios and ontology to express the relationship between cyber assets* (2009). <http://www.dtic.mil/cgiibin/GetTRDoc?AD=ADA517410>
- Buchler, N., Rajivan, P., Marusich, L., Lightner, L., Gonzalez, C.: Sociometrics and observational assessment of teaming and leadership in a cyber security defense competition. *J. Comput. Secur.* **73**, 114–136 (2017)
- Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, May 2017
- Caulkins, B.D., Badillo-Urquiola, K., Bockelman, P., Leis, R.: *Cyber workforce development using a behavioral cybersecurity paradigm*. In: Connelly, C., Brantly, A., Thomson, R., Vanatta, N., Maxwell, P., Thomson, D. (eds.) *International Conference for Cyber Conflict US*. Army Cyber Institute, West Point (2016)
- Champion, M.A., Rajivan, P., Cooke, N.J., Jariwala, S.: *Team-based cyber defense analysis*. In: *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pp. 218–221 (2012)
- Champion, M., Jariwala, S., Ward, P., Cooke, N.J.: *Using cognitive task analysis to investigate the contribution of informational education to developing cyber security expertise*. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **58**(1), 310–314 (2014)

- Chen, J.Y., Barnes, M.J.: Human-agent teaming for multirobot control: a review of human factors issues. *IEEE Trans. Hum. Mach. Inter.* **44**(1), 13–29 (2014)
- Cohen, W.W.: Learning rules that classify e-mail. In: AAAI Spring Symposium on Machine Learning in Information Access, vol. 18, p. 25, March 1996
- Cook, M.: Cyber Acquisition Professionals Need Expertise (But They Don't Necessarily Need to Be Experts). Defense Acquisition University, Fort Belvoir (2014)
- Cooke, N.J., Gorman, J.C., Myers, C.W., Duran, J.L.: Interactive team cognition. *Cogn. Sci.* **37**, 255–285 (2013). <https://doi.org/10.1111/cogs.12009>
- Cooke, N.J., Shope, S.M.: Designing a synthetic task environment. In: *Scaled Worlds: Development, Validation, and Application*, pp. 263–278 (2004)
- Dvorsky, G.: Hackers Have Already Started to Weaponize Artificial Intelligence 11 September 2017. <https://gizmodo.com/hackers-have-already-started-to-weaponize-artificial-in-1797688425>. Accessed 22 Feb 2018
- Emmanuel, Z.: Security experts air concerns over hackers using AI and machine learning for phishing attacks, 5 October 2017. <http://www.computerweekly.com/news/450427653/Security-experts-air-concerns-over-hackers-using-AI-and-machine-learning-for-phishing-attacks>. Accessed 22 Feb 2018
- Franke, U., Brynielsson, J.: Cyber situational awareness: a systematic review of the literature. *Comput. Secur.* **46**, 18–31 (2014)
- Gonzalez, C., Ben-Asher, N., Oltramari, A., Lebiere, C.: Cognitive models of cyber situation awareness and decision making. In: Wang, C., Kott, A., Erbacher, R. (eds.) *Cyber Defense and Situation Awareness*. Springer (in press)
- Granåsen, M., Andersson, D.: Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. *Cogn. Technol. Work* **18**(1), 121–143 (2016)
- Hackman, J.R., Katz, N.: *Group Behavior and Performance*, pp. 1208–1251. Wiley, New York (2010)
- Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., Tygar, J.D.: Adversarial machine learning. In: *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, pp. 43–58. ACM, October 2011
- Internet Live Stats - Internet Usage & Social Media Statistics. Accessed 19 Feb 2018. <http://www.internetlivestats.com/>
- Jajodia, S., Liu, P., Swarup, V., Wang, C.: *Cyber Situational Awareness*. Springer Publishing, New York (2010)
- Jariwala, S., Champion, M., Rajivan, P., Cooke, N.J.: Influence of team communication and coordination on the performance of teams at the iCTF competition. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56(1), pp. 458–462. SAGE Publications, September 2012
- Joint Publication 3–12: *Cyberspace Operations*, Washington, DC: Joint Chiefs of Staff, U.S. Department of Defense (2013)
- Knowles, W., Prince, D., Hutchison, D., Disso, J.F., Jones, K.: A survey of cyber security management in industrial control systems. *Int. J. Crit. Infrastruct. Protections* **9**, 52–80 (2015)
- Kott, A., Alberts, D., Zalman, A., Shakarian, P., Maymi, F., Wang, C., Qu, G.: Visualizing the tactical ground battlefield in the year 2050: Workshop report (No. ARL-SR-0327). Army Research Lab Adelphi Maryland (2015)
- Metz, R.: Why Microsofts teen chatbot, Tay, said lots of awful things online, 24 March 2016. <https://www.technologyreview.com/s/601111/why-microsoft-accidentally-unleashed-a-neo-nazi-sexbot/>. Accessed 23 Feb 2018
- Newhouse, B., Keith, S.S., Witte, G.: *NICE Cybersecurity Workforce Framework*. National Institute of Standards and Technology, Gaithersburg (2016)

- Onwubiko, C., Owens, T.J.: *Situational Awareness in Computer Network Defense: Principles, Methods and Applications*. Information Science Reference, Hershey (2011)
- Paletz, S.B.F.: Multidisciplinary teamwork and big data. In: *Human-Centered Big Data Workshop*, At Raleigh, NC (2014). <https://doi.org/10.1145/2609876.2609884>
- Papernot, N., McDaniel, P., Sinha, A., Wellman, M.: Towards the science of security and privacy in machine learning (2016). arXiv preprint [arXiv:1611.03814](https://arxiv.org/abs/1611.03814)
- Pellerin, C.: *Work: Human-Machine Teaming Represents Defense Technology Future* (2015). <https://www.defense.gov/News/Article/Article/628154/work-human-machine-teaming-represents-defense-technology-future/>. Accessed 1 Feb 2018
- Pfleeger, S.L., Caputo, D.D.: Leveraging behavioral science to mitigate cyber security risk. *Comput. Secur.* **31**(4), 597–611 (2012)
- Proctor, R.W., Chen, J.: The role of human factors/ergonomics in the science of security decision making and action selection in cyberspace. *Hum. Factors J. Hum. Factors Ergon. Soc.* (2015). <https://doi.org/10.1177/0018720815585906>
- Rajivan, P., Champion, M., Cooke, Nancy J., Jariwala, S., Dube, G., Buchanan, V.: Effects of teamwork versus group work on signal detection in cyber defense teams. In: Schmorrow, Dylan D., Fidopiastis, Cali M. (eds.) *AC 2013. LNCS (LNAI)*, vol. 8027, pp. 172–180. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39454-6_18
- Rajivan, P.: *Information Pooling Bias in Collaborative Cyber Forensics* (Doctoral dissertation, Arizona State University) (2014)
- Rajivan, P., Cooke, N.: *Information Pooling Bias in Collaborative Security Incident Analysis*. Human Factors (in press)
- Rajivan, P., Moriano, P., Kelley, T., Camp, J.: Factors in an end user security expertise instrument. *Inf. Comput. Secur.* **25**(2), 190–205 (2017)
- Salas, E., Dickinson, T.L., Converse, S.A., Tannenbaum, S.I.: Toward an understanding of team performance and training. In: *Teams their Training and Performance*, pp. 3–29 (1992)
- Seymour, J., Tully, P.: *Weaponizing data science for social engineering: Automated E2E Spear Phishing on Twitter*. Black Hat USA, 37 (2016)
- Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, May 2017
- Shortliffe, E.: *Computer-Based Medical Consultations: MYCIN*. Elsevier, New York (2012)
- Simsarian Webber, S.: Leadership and trust facilitating cross-functional team success. *J. Manage. Dev.* **21**(3), 201–214 (2002)
- Spiro, R.J.: *Cognitive Flexibility Theory: Advanced Knowledge Acquisition in Ill-Structured Domains*. Technical Report No. 441 (1988). <http://eric.ed.gov/?id=ED302821>. Accessed 5 Oct 2017
- Srinidhi, B., Yan, J., Tayi, G.K.: Allocation of resources to cyber-security: the effect of misalignment of interest between managers and investors. *Decis. Support Syst.* **75**(1), 49–62 (2015). <http://doi.org/10.1016/j.dss.2015.04.011>
- Stasser, G., Titus, W.: Pooling of unshared information in group decision making: biased information sampling during discussion. *J. Pers. Soc. Psychol.* **48**(6), 1467 (1985)
- Trexler, E.: *Why Human-Machine teaming is the future of cybersecurity* (2017). <https://federalnewsradio.com/commentary/2017/11/why-human-machine-teaming-is-the-future-of-cybersecurity/>. Accessed 1 Feb 2018
- Tung, L.: IoT devices will outnumber the world's population this year for the first time 13 February 2017. <http://www.zdnet.com/article/iot-devices-will-outnumber-the-worlds-population-this-year-for-the-first-time/>. Accessed 19 Feb 2018

- Veksler, B.Z.: Visual search strategies and the layout of the display. In: Salvucci, D.D., Gunzelmann, G. (eds.) *Proceedings of the 10th International Conference on Cognitive Modeling*, pp. 323–324. Drexel University, Philadelphia (2010)
- Vicane, A., Funke, G., Mancuso, V., Greenlee, E., Dye, G., Borghetti, B., Brown, R.: Coordinated displays to assist cyber defenders. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60(1), pp. 344–348. SAGE Publications, September 2016
- Whitley, J., D’Amico, K.: The real work of computer network defense analysts. In: Goodall, J.R., Conti, G., Ma, K.L. (eds.) *Workshop on Visualization for Computer Security*, pp. 19–37 (2008)



Automatically Unaware: Using Data Analytics to Detect Physiological Markers of Cybercrime

Nancy Mogire¹(✉), Randall K. Minas², and Martha E. Crosby¹

¹ Information and Computer Sciences, University of Hawaii at Manoa, POST
317 1680 East-West Road, Honolulu, HI 96822, USA

nmogire@hawaii.edu

² Shidler College of Business, University of Hawaii at Manoa, 2404 Maile Way
Suite E601f, Honolulu, HI 96822, USA

rminas@hawaii.edu

Abstract. Cybercrime investigation is reliant on availability of adequate and valid digital artifacts useable for reconstructing security incidents or triangulating other available information to make it useful. Various operational artifacts of computer systems, networks and software have been studied and gradually applied as forensic evidence. However the scope of studies on human-generated artifacts as forensic evidence has been limited mostly focusing on surveillance images, with DNA deposits being widely studied via older forensic fields. We present the case that further focus on human-centric evidence in form of physiological measurements is useful in triangulating other evidence as well as in making some direct inferences. In this concept paper: we pair electroencephalography (EEG) with change point detection algorithms to conceptually model the acquisition and processing of EEG signals into forensic artifacts; propose continuous data reduction and packaging to keep the system forensic-ready; suggest a schema for validating such artifacts towards their applicability as forensic evidence; and model a study to be used in testing the conceptual model. This work contributes to cybersecurity research by highlighting human-generated artifacts as a forensic big data resource and presenting a methodology for harnessing the data to turn it into useful information.

Keywords: Digital forensics · Forensics artifacts · Physiological measures
Electroencephalography · Cybersecurity

1 Introduction

CYBERCRIME, crimes that are committed electronically, take various forms. Often, the computer is either a tool or a target. The two aspects have sometimes been categorized broadly as computer-enabled and computer-dependent crimes [1]. The increase in cybercrime over the past decades has been staggering, including increased financial scams, child pornography, illegal gambling, cyber bullying, espionage, system vandalism using viruses and election interference [2]. The proliferation of cybercrime over the years led to the growth of computer forensics as a new forensic field.

Forensics involves acquisition and analysis of relevant data for application as evidence in legal or civil cases. There are many informational artifacts in networks

including network activity metadata, system status and usage details as well as software activity data. However, there is also seemingly untapped data from the human component of the computing system. Some of this has been utilized (e.g., DNA and camera footage), yet the power of psychophysiological data of individuals has remained largely unexplored in forensics.

There are many forms of psychophysiological data, including skin conductance, heart rate, and eye movement data. These forms of psychophysiology have been used in the forensics field for years, but have not been used extensively to detect computer crimes. In addition, electroencephalography (EEG) provides electrical data present at the scalp that elucidates aspects of cognition and emotion. These physiological measures have been unharnessed in their ability to provide large data sets through which forensics of computer-based crimes might be detected. The power of psychophysiological data is that it is difficult and often, impossible for the user to disguise their behavior as it is preconscious.

Building off Kahneman's research [3]. System 1 is the unconscious automatic cognition that is immediate and reactive, making decisions in less than a second, and System 2 is the deliberate, thoughtful process found in theories. Most researchers conclude that much human behavior is driven by System 1, with some believing that most behavior is controlled by System 1 [3, 4]. The exact amount of behavior controlled by System 1 is debatable and lies beyond the scope of this paper. However, the key point from this research is that most researchers studying dual process cognition would agree that a meaningful amount of behavior is controlled by System 1 cognition. The key takeaway is that the cybercriminal is unable to readily hide their System 1 behavior as much of it is unconscious. This provides a valuable pathway for data analytics to be combined with psychophysiological measurement to identify events related to cybercrime.

2 Background and Related Work

2.1 Digital Forensics

Digital forensics is a term often used interchangeably with computer forensics but that has evolved to cover investigation of all kinds of devices capable of storing data [5]. The term has been defined as the science of locating, extracting, and analyzing of data from different devices, for use by specialists who interpret such data for application as legal evidence [5]. The end goal is to be able to use such evidence in legal testimony that is admissible in a court of law. Such data can also be used in internal corporate investigations. Applications of evidence include attribution of actions towards suspected sources, confirming of alibis or statements, determination of intent and authentication of documents [5]. One of the key considerations in digital forensics is how to collect evidence without altering its contents so as to sustain its admissibility in court. various methods of collecting digital evidence and the attendant issues are discussed in [6].

2.2 Examples of Digital Evidence

System based evidence: This included both volatile or non-volatile data. Volatile data includes system date and time, open TCP and UDP ports and logged on users. Non-volatile data includes operating system version, user accounts and the auditing policy [6].

Network artefacts: can also be used as digital evidence. These include session information, alerts, content transmitted over networks such headers and payload in a data packet as well as network statistics e.g. how many packets are transmitted, source and destination [6].

User information stored in devices: including wearable computing, contains various artifacts such as: Paired device information, voice commands, Bluetooth packet data, text message notifications and recent tasks [7]. The authors note that some of this data requires root access to the device in order to acquire it, which in turn causes a factory reset on the device. With a factory reset, the data may not be admissible as evidence.

2.3 Evidence Acquisition

Forensic investigation can be summarized in five steps: Collection of evidence, preservation and transportation to analysis site, Identification of evidence and planning of processing, analysis of evidence, presentation via reports [5]. The focus of our work relates to the collection or acquisition stage.

Collection of evidence often involves seizure, imaging and analysis of digital media [5]. Neuner et al. [8] discuss one of the major problems of digital forensic data collection process is the need for redundancy. As they discuss, NIST SP-800-86 recommends the use of a working copy and retention of a backup copy in case data becomes tainted during analysis. These needs increase both time and storage space overhead. The solution proposed is the reduction of a working copy by removing unnecessary and duplicate files.

2.4 Anti-forensics

Anti-forensics has been defined as “Attempts to negatively affect the existence, amount, and/or quality of evidence from a crime scene, or make the examination of evidence difficult or impossible to conduct” [9]. There are several anti-forensic techniques including physical destruction of devices, tampering with log files, data hiding, signature masking. However, as discussed in [5], many of them require more advanced computing knowledge. They also present various suggestions of mitigations for these anti-forensic techniques. As they discuss, another redeeming factor is that the lack of evidence is evidence itself provided the investigators can detect the absence of data that should be present. The nature of artifacts determines whether it easy or difficult to detect an absence.

2.5 Digital Forensic Readiness of a System

As Endicott-Popovsky [10] discusses, when incidents happen in networks, the organization can find itself in the dilemma between preserving the state for forensic acquisition and quickly restoring the network. If the organization is not ready for forensic data acquisition then the time and cost may be too great. In that case, the organization may choose to focus on restoring the network which may could diminish the forensic value of relevant files.

Carrier and Spafford [11] present a model for forensic readiness. Their model consists of three phases: the first one is initial readiness which entails setting up of forensic logging, developing and testing tools. The next one is the deployment which entails detection of incidents followed by systematic verification which may lead to the opening of a crime scene investigation if necessary. Following that is investigation which involves searching for evidence and reconstructing events. The final step is the presentation of the evidence and documentation.

Kazadi and Jazri [12] present another version of readiness framework: Their framework begins with risk assessment to identify the most critical data points; development of security policy for protection of software tools used in collecting evidence data and setting of quality standards regarding such tools e.g. audit schedules. The readiness phase then continues to selection of data collection software tools and other equipment such as surveillance cameras and biometric devices. The second phase involves operations decisions such as the sensor placements based on the previously assessed risks. After this point, the data is logged based on predefined rules, after which it is preserved and stored in case it is needed. The stored logs are later processed for deletion based on a preset time policy. The deletion phase involves analysis and report generation before authorized deletion.

2.6 Validating Forensic Artifacts

Palmer [13] discusses that validity of evidence must be ascertained before it can be considered judicious and reliable in a legal argument. They further discuss that validating digital evidence requires verification of the relevant parts of such evidence including how it is created, processed, stored and transferred, so as to establish a confidence level and hence a value for the inferences drawn from such evidence. If validity of such evidence cannot be established then its weight can be at best diminished and at worst negated.

Factors that can affect validity of evidence include; evidence taken out of context or misinterpreted, misleading or false evidence, failure to identify relevant errors and difficulty in reconstructing the evidence when questions are raised during the validation process [13, 14]. Boddington et al. [15] discuss that evidence validation requires a chain of proof assertions made for such evidence. For each assertion made for the validity of the evidence, a stronger negative assertion may exist in which case such evidence may be weakened. The authors give an example of an assertion *<<data was overwritten by virus scanner>>* which gets negated by: *<<file creation date is unchanged>>*. An investigator can use a checklist that is artifact-specific to run a piece

of forensic data through an assertion chain and thereby make a claim to its validity or invalidity.

2.7 Physiological Artifacts for Digital Forensics

Mobile phone swipe gestures have been considered as a source of forensic data by Mondal and Bours [16] who employ a continuous identification [CI] model to continually verify the user. They suggest that in a closed user group, apart from detecting that an unauthorised user is on the phone and locking it, the system can also attempt to identify the adversary and store the information for evidential use. They note that this may be the first application of biometric information for continuous identification [CI]. The forensic angle of their work is that they obtain the identity of the intruder and store it as evidence. As they discuss, their focus is the protection of the system which entails locking the phone as soon as an unauthorized user is detected. The forensic extent of their work is that they obtain the identity of the intruder and store it as evidence although accuracy of this process as they note is hindered by the focus on quickly locking out the impostor [16].

Other collectable physiological measurements could potentially be applicable as digital forensic artifacts. These include heart rate, skin conductance, breath count, and electroencephalograms. As an initial focus, in the next section we consider electroencephalograms for their potential applicability.

2.8 Suitability EEG as a Source of Forensic Data

Electroencephalography (also abbreviated as EEG) is the recording of the brain's electric potentials varying in time at different frequencies per second and range from a few microvolts to a few millivolts [17]. Electroencephalograms are produced as a result of the synchronous action of numerous neurons in the brain [18]. EEG measurements are recorded from scalp electrodes [19]. Nunez and Katznelson [17] discuss that EEG signals occur and are recordable without the need for any deliberate stimuli application. For the information to be applicable for a desired usage, relevant stimuli is applied and the changes are recorded as a measure of the response to the stimuli. These changes are referred to as evoked potentials [17].

EEG signals provide a unique signature and can be used to distinguish people [18, 20]. They are collectable and quantifiable by amplitude or energy changes and hence the brain's responses to stimuli can be calibrated and measured [20]. EEG signal acquisition is applicable to most people since different forms of stimuli can trigger response signals and this accommodates for different impairments. Stimuli forms include visual, somatosensory and motor imagery [21].

Examples of media used to present visual stimulation include the Snodgrass and Vanderwart picture set which induces highly synchronized neural activity in the gamma band [22]. The Snodgrass and Vanderwart picture set is a standard set of 260 pictures drawn in black lines based on a set of rules that provide consistency in pictorial representations [20]. Lists of acronyms have also been used as stimuli for Visual Evoked Potentials [21].

EEG has been utilized for brain-computer interfaces (BCI) in medical and non-medical settings. Several applications have been derived including: automated diagnosis of epileptic EEG using entropies [24]; automated drowsiness detection using wavelet packet analysis [25]; EEG-based mild depression detection using feature selection methods and classifiers [26]; neuro-signal based lie detection [27]; authentication [20] and continuous authentication.

Several different methods have been applied for analysis of the data. Many but not all of these methods involved are machine learning based. In authentication applications, the analysis relies on a data training phase before the trained model can be used for authentication.

Model training methodology does not lend itself to flexible applicability in a forensic data collection setup. As Al Solami et al. [28] discuss, there are various limitations of these training model based schemes that makes them impractical including: biometric data is not always available in advance of events; the need for too many training samples can be impractical; behavior biometric can change between the training and testing phase; behavior biometrics can change from one context to another.

2.9 Change Point Detection in EEG Data

Al Solami et al. [28] propose a generic model of application of a change-point detection in the case of continuous authentication systems, to remove the need for prior data training and potentially increase speed. This technique has been used in detecting change points in various multivariate time series. Yu et al. [29] applied this method for detecting changes in statistical dependence within the time series, modelled using Gaussian copula graphical models. The data between the changepoints is regarded as piecewise stationary. Among various applications, the authors analyzed EEG using change point detection for the medical application of predicting epileptic seizures. During the recording of the data, the patients had experienced events that were judged to be clinical seizures by experts. Using a change point detection method two change points are observed at two seconds before the start and the end of the seizure, which is in line with the clinical specialists initial expectations.

Change point detection has also been used in conjunction with a smart home activity recognition sensor data for detecting activity transitions, segmenting the activities into separate actions and correctly identifying each action. The intended applications include timing notifications and interventions [30]. Their work involves a data training phase so as to provide activity labels needed in their model.

The non-model training property of change point detection methodologies makes the data defensible as a forensic artifact as it reduces biased manipulation and chances of distortion. The data can be used to map the onset of irregular and perhaps critical cognitive events on the part of the user, which may also signal the change of signal source.

There are several change detection algorithms with various strengths and weaknesses. The method particularly amenable to real time processing of EEG data is the cumulative sum (CUSUM) algorithm. Below is the general form of a change detection algorithm:

```

initialization
  if necessary
end
while the algorithm is not stopped do
  measure the current sample  $x[k]$ 
  decide between  $H_0$  (no change) and  $H_1$  (one change)
  if  $H_1$  decided then
    store the detection time  $nd \leftarrow k$ 
    estimate the change time  $nc$ 
    stop or reset the algorithm
  end
end

```

Algorithm Source: Granjon [31]

The algorithm has two key components namely: **detection**: a decision between change and no change at each step; and **estimation**: of when the change occurred. The details of setup of both components are crucial from a forensic perspective, in order to reliably triangulate the onset of interesting events in or around a system.

3 Proposed Methodology

Several parts are considered in modelling the framework for EEG- based forensic evidence

- I. Some forensic features of EEG and their contextual applicability as forensic artifacts
- II. Forensic readiness plan for building EEG signal evidence
- III. Process flow for acquisition, analysis and storage of EEG forensic data
- IV. The change point detection algorithm
- V. Data packaging and reduction - to reduce data management overhead
- VI. Eeg evidence validation schema
- VII. Other issues
 - A. Anticipating Anti-forensics
 - B. Privacy and Ethics Argument

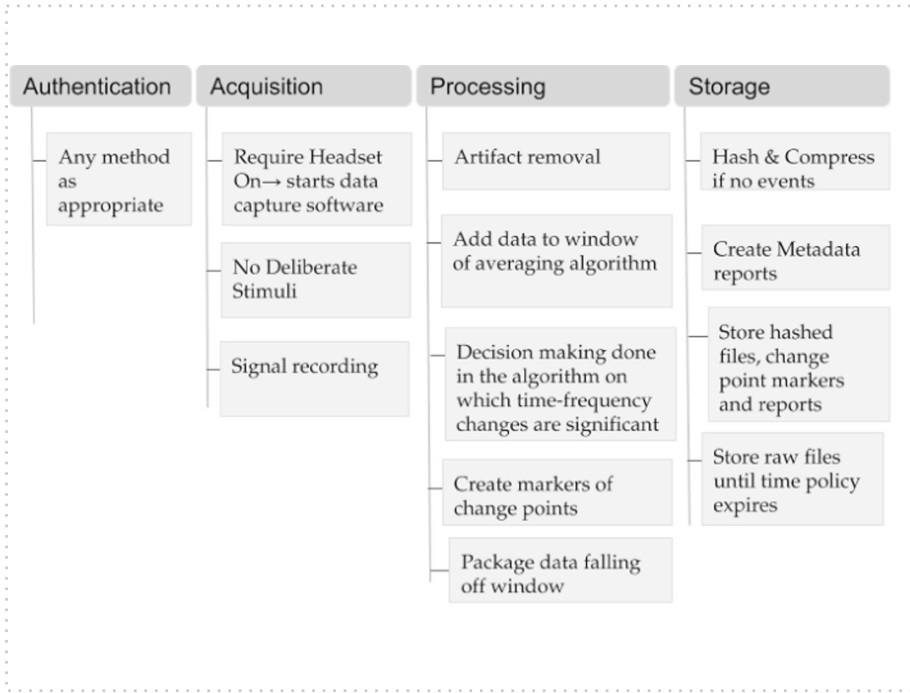
I. Some Features of EEG and Their Contextual Applicability as Forensic Artifacts

- A. **Uniqueness to Individual:** EEG Signals have unique features that can be used to distinguish between individuals. This property can be applicable if the forensic investigation is concerned with verifying that an intruder was logged into a system.
- B. **Can show responses to stimuli:** The changes caused by exposure to stimuli are known as evoked potentials. These can be applicable if an investigation is concerned with verifying recognition or recall of some stimuli being shown to the subject.
 Response to stimuli can also be applicable if the investigation is concerned with finding out when a unusual activity may have began around the vicinity of the person and device under investigation.
- C. **EEG can show cognitive deviations:** There are five primary EEG signal bands classified by frequency. Other than that, EEG has been used to show changes in state of mind e.g. epileptic seizures, fatigue, and drowsiness as discussed earlier. This property can be applicable if an investigation is concerned with whether a person’s state of cognition deviated abnormally during performance if a critical task e.g. operating critical equipment.

II. Forensic Readiness Plan for Building EEG Signal Evidence

Readiness Phase:	Deployment Phase:	Evidence Application or Storage:
<ul style="list-style-type: none"> - Identify parts of system usage where EEG data collection can be meaningful Select: <ul style="list-style-type: none"> - Headset Type - No. of channels - Data capture software 	<ul style="list-style-type: none"> - Collect signal - Process signal with change detection module - Package signal - Build metadata, alerts etc.. - Update reports - Hash, compress and store signal as per time policy 	<ul style="list-style-type: none"> - If incident is found, provide data to forensic examiner - If no incident found, retain per time policy and trigger disposal procedure Disposal <ul style="list-style-type: none"> - Analyze data - Summarize - Authorized disposal

III. Process Flow Diagram for Acquisition, Analysis and Storage of EEG Forensic Artifacts



IV. Change Point Detection

There is an initial period when no averaging is done until the minimum window size is reached. The algorithm is applied with an appropriate window size to the EEG data. The EEG data collected in the sample can be decomposed and analyzed using Independent Components Analysis (ICA). A common problem in neuroimaging research results from the collection of large amounts of data which, based upon the Central Limit Theorem, become normally distributed. However, the brain is comprised of discrete patches of cortex that are very active at some points in time and relatively inactive at others (i.e., activity is not normally distributed across the scalp) [32]. ICA overcomes this problem by taking this Gaussian data and rotating it until it becomes non-Gaussian, thereby isolating independent components contributing to the activation.

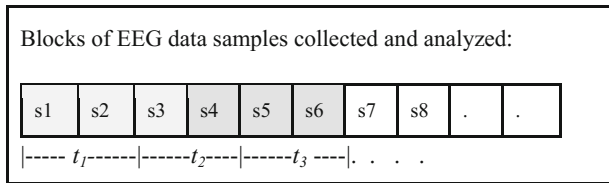
Initially, an EEGLab ICA performs a Principal Components Analysis (PCA). At each electrode site the program assesses which of the other electrode sites account for the most variance in the signal. Taking these weighted values it then relaxes the orthogonality constraint of PCA to isolate individual components of activation [32]. Each ICA component then represents a pattern of activation over the entire brain, not solely the activity present at a specific electrode. The number of independent components (ICs) depends on the number of electrodes in the dataset, as the algorithm is working in an N-dimensional space (where N is the number of electrodes). Therefore an n-channel EEG system would produce an ICA equation similar to:

$$IC_1 = w_1e_1 + w_2e_2 + \dots w_n e_n$$

Where IC is the ‘‘Independent Component’’, w is the weight assigned to the electrode, and e is the electrode. The change detection algorithm that will be used for pattern detection will take in one IC as an input at each point.

V. Data Reduction and Packaging

One of digital forensic problems is the overhead of handling unwieldy volumes of data [8]. Further, processing data when it is already in duplicates, scattered and tampered with is more likely to result in invalid evidence. Data packaging and reduction is aimed at organizing data early on within the framework of a forensic ready system. Putting data into organized formats can support fast and quality analyses later if incidents happen. In this model, organization of data consists of incremental categorization, analysis, decisions, metadata, reports, hashing, compression and finally storage.



Using a time variable package blocks of data that have fallen off the left side of the change detection analysis window. After every t seconds unless event, if event, process event details first then package data. Packaging involves hashing and compression followed by labelling and storage.

VI. Evidence Validation Schema

Validity has been defined as: ‘‘the extent to which a concept, conclusion or measurement is well-founded and corresponds accurately to the real world.... validity of a measurement tool is considered to be the degree to which the tool measures what it claims to measure; in this case, the validity is an equivalent to accuracy.’’ [33].

Boddington et al. [15] discuss that validation of digital evidence requires verifying the path of the evidence since its creation i.e. the digital environment in which it was created, processed and transferred, including the evidence file itself plus the software and hardware used in its handling. As they further discuss, If validity of evidence cannot be established then its weight is diminished or even negated.

This representation allows forensic examiner to list the relevant validity properties to check, and later add what the defending team may bring up.

	Assertion	Verification [+1]	Negation [-1]	Assertion weight
Start				
A_1				

(continued)

(continued)

	Assertion	Verification [+1]	Negation [-1]	Assertion weight
A ₂				
.				
.				
A _m				
End				Total weight

Schema derived from: Boddington et al. [15]

4 Planned Study: Applicability of EEG Data for Confirming Events

In this study we plan to test the usability of EEG for triangulating the time that an event may have occurred on or around a computing system. The signal acquisition begins with general activities for the participant to use the device. At irregular but predefined intervals, we introduce a predetermined and timed cyber event. Example events include: constructing a cyber event on the participant's path by including a privilege escalation opportunity as follows: the participants will be asked to log into the test application having created an account. When they click the login link, in the form that comes up they will find the name "admin" pre-filled with a masked out password. It will look like an accidental flaw. Ideally they should erase and input their own credentials which will take them to their own account normally. However, they would be in a position to simply click enter on the admin credentials and attempt to go into admin space. The signal continues to be collected during the event intervals. We will investigate whether the signal once acquired and tested by the algorithm will be able to reveal that the events introduced occurred at given points.

5 Future Work

In this paper we have focused on turning EEG artifacts into forensic evidence. However, other physiological measurements can be similarly applicable. Some work would need to be done to determine which how the various physiological measurements can be harnessed into forensic evidence and how such models could be tested. Another questions that could be answered by further work is how to deal with anti-forensics activities such as rogue physiological data insertions and deletions as well as the creation of deliberate physiological noise during system usage where physiological data is collected.

6 Discussion and Conclusion

Many problems in cybersecurity involve individuals. Since cognition often occurs in the automatic System 1 realm, where they act in automatic mode and are unaware of specific interactions with the system. The human's system 1 or automatic behavior dimension has been largely neglected and it presents an opportunity to be applied to forensic-readiness in digital systems. Physiological responses to events can be captured and analyzed to help detect events, and when events occur, previously processed and packaged data can be readily available to facilitate the forensic investigation process. This application is especially relevant for critical systems which may include government sensitive data systems or SCADA systems which are at high risk from both internal and external threats.

References

1. Cybercrime: Legal Guidance: Crown Prosecution Service. Cps.gov.uk (2017). http://www.cps.gov.uk/legal/a_to_c/cybercrime/#a03. Accessed 28 Oct 2017
2. Holt, T., Bossler, A.: An assessment of the current state of cybercrime scholarship. *Deviant Behav.* **35**(1), 20–40 (2013)
3. Kahneman, D.: *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York (2015)
4. Stanovich, K., West, R.: Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* **23**(5), 645–665 (2000)
5. Hausknecht, K., Gruicic, S.: Anti-computer forensics. In: 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (2017)
6. Resendez, I., Martinez, P., Abraham, J.: *An introduction to digital forensics* (2017)
7. Baggili, I., Oduro, J., Anthony, K., Breiting, F., McGee, G.: Watch what you wear: preliminary forensic analysis of smart watches. In: 2015 10th International Conference on Availability, Reliability and Security (2015)
8. Neuner, S., Mulazzani, M., Schrittwieser, S., Weippl, E.: Gradually improving the forensic process. In: 2015 10th International Conference on Availability, Reliability and Security (2015)
9. Harris, R.: Arriving at an anti-forensics consensus: examining how to define and control the anti-forensics problem. *Digit. Invest.* **3**, 44–49 (2006)
10. Endicott-Popovsky, B.: *Digital evidence and forensic readiness* (2017)
11. Carrier, B., Spafford, E.: An event-based digital forensic investigation framework (2004). http://www.digital-evidence.org/papers/dfrws_event.pdf. Accessed 28 Oct 2017
12. Kazadi, J., Jazri, H.: Using digital forensic readiness model to increase the forensic readiness of a computer system. In: 2015 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC) (2015)
13. Palmer, G.: Forensic analysis in a digital world. *Int. J. Digit. Evid.* **1**(1), 1–6 (2002)
14. Cohen, F.: *Challenges to Digital Forensic Evidence*, 129 p. Fred Cohen & Associates, Livermore (2008). ISBN 1-878109-41-3
15. Boddington, R., Hobbs, V., Mann, G.: Validating digital evidence for legal argument. In: Australian Digital Forensics Conference (2017)

16. Mondal, S., Bours, P.: Continuous authentication and identification for mobile devices: combining security and forensics. In: 2015 IEEE International Workshop on Information Forensics and Security (WIFS) (2015)
17. Nunez, P., Katznelson, R.: *Electric Fields of the Brain*. Oxford University Press, New York (1981)
18. Başar, E.: *Brain Function and Oscillations*. Springer, Heidelberg (1998). <https://doi.org/10.1007/978-3-642-72192-2>
19. Regan, D.: *Human Brain Electrophysiology*, pp. 1–147. Elsevier, New York (1989)
20. Zuquete, A., Quintela, B., Cunha, J.: Biometric authentication using electroencephalograms: a practical study using visual evoked potentials. *Electrónica e Telecomunicações* **5**(2), 185–194 (2010)
21. Palaniappan, R.: Electroencephalogram-based Brain–Computer Interface: an introduction. In: Miranda, E.R., Castet, J. (eds.) *Guide to Brain–Computer Music Interfacing*, pp. 29–41. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6584-2_2
22. Snodgrass, J., Vanderwart, M.: A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *J. Exp. Psychol. Hum. Learn. Memory* **6**(2), 174–215 (1980)
23. Gui, Q., Jin, Z., Xu, W.: Exploring EEG-based biometrics for user identification and authentication. In: 2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) (2014)
24. Acharya, U., Molinari, F., Sree, S., Chattopadhyay, S., Ng, K., Suri, J.: Automated diagnosis of epileptic EEG using entropies. *Biomed. Signal Process. Control* **7**(4), 401–408 (2012)
25. da Silveira, T., Kozakevicius, A., Rodrigues, C.: Automated drowsiness detection through wavelet packet analysis of a single EEG channel. *Expert Syst. Appl.* **55**, 559–565 (2016)
26. Li, X., Hu, B., Sun, S., Cai, H.: EEG-based mild depressive detection using feature selection methods and classifiers. *Comput. Methods Programs Biomed.* **136**, 151–161 (2016)
27. Cakmak, R., Zeki, A.: Neuro signal based lie detection. In: 2015 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS) (2015)
28. Al Solami, E., Boyd, C., Clark, A., Islam, A.: Continuous biometric authentication: can it be more practical? In: 2010 IEEE 12th International Conference on High Performance Computing and Communications (HPCC) (2010)
29. Yu, H., Li, C., Dauwels, J.: Network inference and change point detection for piecewise-stationary time series. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2014)
30. Aminikhanghahi, S., Cook, D.: Using change point detection to automate daily activity segmentation. In: 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops) (2017)
31. Granjon, P.: The CUSUM algorithm a small review (2012). http://chamilo2.grenet.fr/inp/courses/ENSE3A35EMIAAZ0/document/change_detection.pdf. Accessed 28 Oct 2017
32. Onton, J., Westerfield, M., Townsend, J., Makeig, S.: Imaging human EEG dynamics using independent component analysis. *Neurosci. Biobehav. Rev.* **30**(6), 808–822 (2006)
33. Validity (statistics): *En.wikipedia.org* (2017). [https://en.wikipedia.org/wiki/Validity_\(statistics\)](https://en.wikipedia.org/wiki/Validity_(statistics)). Accessed 28 Oct 2017



Understanding Behaviors in Different Domains: The Role of Machine Learning Techniques and Network Science

Grace Teo¹(✉), Lauren Reinerman-Jones¹, Joseph McDonnell², Hayden J. Trainor³, Rainier A. Porras³, and Jacob G. Feuerman³

¹ Institute for Simulation and Training, University of Central Florida, Orlando, FL, USA

{gteo, lreiner}@ist.ucf.edu

² Dynamic Animation Systems, Fairfax, VA, USA

joe.mcdonnell@d-a-s.com

³ United States Military Academy, West Point, NY, USA

{hayden.trainor, rainier.porras, jacob.feuerman}@usma.edu

Abstract. Recent developments in the Internet of Things (IoT), social media, and the data sciences have resulted in larger volumes of data than ever before, offering more opportunity for observing and understanding behaviors. Advances in data analytic and machine learning techniques have also enabled assessments to be more multi-faceted, incorporating data from more sources. Machine learning algorithms such as Decision Trees and Random Forests, K-nearest neighbors, and Artificial Neural Networks have been used to uncover hidden patterns in data and derive predictions and recommendations from a wide range of data types and sources. However, these do not necessarily yield insights into behaviors in complex systems/domains. Methods from mathematics such as Set Theory, Graph Theory, and Network Science may be useful in shedding light on the interactions and relationships within and across domains. This paper provides a description of the applications, strengths, and limitations of some of these techniques and methods.

Keywords: Machine learning techniques · Decision tree · Random forest · K-nearest neighbor · Artificial Neural Network · Network science

1 Introduction

Most of the data in the world today has only been created within the last few years [1], and the amount of data generated daily is projected to only increase, especially with the rise of the Internet of Things (IoT) and social media. One report projected that by 2020, each individual would create about 1.7 megabytes of new information every second [2]. All this offers unprecedented opportunities for assessments to understand various behavioral phenomena. With this growth of big data, there has also been a surge in the number of data analytic techniques. Some of these techniques employ machine learning, which has been applied to analytic problems such as prediction, classification,

clustering, and revealing associations. These can be helpful in addressing research questions such as:

- What task or person characteristics predict trust in automation? (Prediction)
- What indicators tend to cluster and do they suggest a new construct, e.g., the construct of *fitness for duty* from hours of sleep, blood alcohol level, cardiovascular functioning [3, 4]? (Clustering)
- How to classify someone as being in high vs. low workload? (Classification)

2 Machine Learning

Machine learning techniques can be categorized into those where the machine is trained with data consisting of inputs with the corresponding behavioral outcomes (supervised learning), and those where the behavioral outcome is unknown and the machine is simply tasked to uncover hidden patterns or structure in the data (unsupervised learning). Unlike unsupervised techniques which have limited applications, supervised learning techniques are more commonplace [5]. They can be used to identify precedents of certain behaviors. Examples of supervised learning techniques include decision trees and random forests, k-nearest neighbor, and artificial neural networks. The data that the machine uses to learn or is trained on, is called *training data*. The new data to which the predictive, machine-developed algorithm or model is applied, is *test data*.

3 Decision Trees and Random Forests

Decision trees are a supervised, predictive modeling technique where input variables are expressed as decision rules that are applied in succession (i.e., recursive partitioning of data) with the goal of classifying observations into their outcomes classes. In doing so, the most influential inputs that relate to the outcome are identified. This result can then be used to predict the outcome class for new observations. Decision trees can be constructed by numerous software programs, both basic and sophisticated such as Microsoft Excel, Weka, etc. Here we will examine how simple decision trees and random forests are constructed, how they can be applied to data sets and what their strengths and limitations are.

3.1 Use of Decision Trees

A decision tree comprises a set of decision rules that are used to classify observations into their outcome classes by the values of their inputs. They can also be used to predict the outcome class of a new observation. Following the creation of a decision tree, the new observation data is run through the tree, which functions like a flowchart, where decision rules “fork” observations into different branches. In the decision tree, the decision rules are the “nodes” that determine the branch that the observations falls into. This branching at the nodes occurs in succession until the observation is eventually

classified into an outcome class at the leaf level. The inputs that specify the decision rules comprise the algorithm of the decision tree.

Figure 1 depicts a decision tree showing the algorithm in predicting a new customer's decision to purchase a cell phone with the newest technology. In this hypothetical decision tree, the observations can be fully classified (i.e., each end node is homogeneous) by the inputs of age, condition of existing phone, and presence of current sales promotions. This means that if all these three pieces of information are known about a new observation, the behavioral outcome of the new customer can be predicted. That is, a new customer will purchase the phone with the latest technology if (i) s/he is below 50 years old and has a phone in poor condition, or (ii) if s/he is below 50 years old and is offered a sale promotion when his/her phone is in good condition.

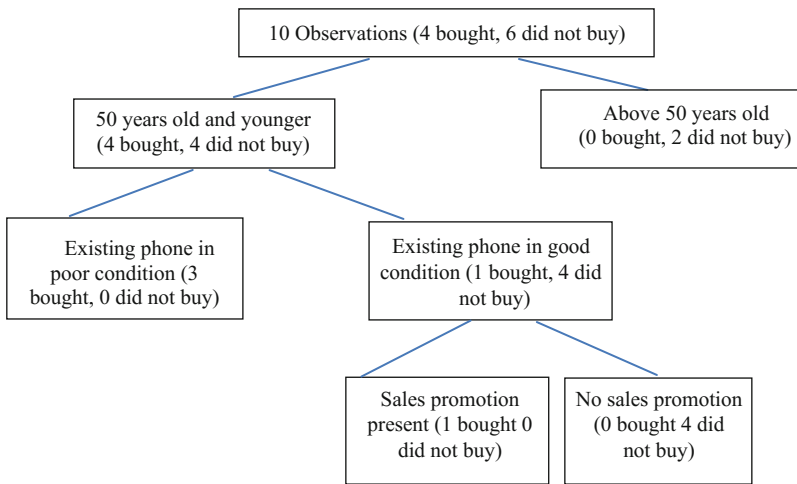


Fig. 1. Example of a decision tree

3.2 Constructing a Decision Tree

In constructing a decision tree, a greedy splitting approach is usually adopted. This process examines the input variables and chooses a tree path that minimizes a particular cost function. Classification utilizes the Gini cost function, while regression (i.e., regression trees) utilizes the sum of squared errors (SSE) cost function [5]. Once a basic decision tree is constructed, it is typically pruned in order to prevent overfitting of the data. Overfitting occurs when the model accurately assesses the training data but fails to accurately assess test data. Pruning is a means to combat this problem by removing each leaf node one by one, evaluating the effect on the cost function after each removal. There are many means to prune a decision tree, but a typical rule of thumb is that the smaller the decision tree, the less likely the overfit, and the more likely it is to be successful with the test data.

3.3 Random Forests

Random forests, or random decision trees, are ensemble methods involving multiple decision trees. Random forests essentially combine multiple decision trees to create multiple classifiers or predictors. The data will then be classified by the mode of these decision trees. Random forests are able to overcome error associated with using only one decision tree by establishing randomness across multiple trees. In creating a random forest, each decision tree uses a random selection of data and a random selection of variables [6]. This allows the trees in a random forest to examine a subset of the data while not focusing on all of the training data. The random forest examines all of the training data by utilizing numerous random decision trees. This method permits overlap among the trees but also prevents the classification or prediction to be made solely by one decision tree.

3.4 Data Required

Decision trees and random forests can be built from various types of data. Large data sets can be easily classified by a larger tree that has numerous nodes or splitting points [5]. Not all decision trees have binary splits; any number of splits may occur. For instance, one variable labeled “age” may split data into “50 years old or younger” and “Above 50 years old,” or be characterized as having four age classes from “0–20 years old,” “21–40 years old,” “41–60 years old,” and “Above 60 years old.” Decision trees do not require the dataset to be complete and can also be constructed if there are missing data for some observations [7]. Due to the random selection of data and variables of random forests, if a variable in a data set is omitted, some decision trees within the random forest may not execute while the random forest as a whole will still produce a reasonable classification or prediction.

3.5 Strengths

One of the most beneficial attributes of decision trees is that they can be easily interpreted through graphics. Although this may become difficult with larger trees, smaller trees can be easily described through a simple diagram and explanation [5, 8, 9]. Curram [9] even explains that this may contribute to insight into factor relationships. Another positive attribute of decision trees is that data does not need to be transformed in any way, as there are no assumptions about the normality of the data or the underlying distribution of the data. Nodes within a decision tree can evaluate both quantitative and qualitative measurements without transforming one into the other. Due to the randomness of a random forest, any inaccuracy of one decision tree can be overcome by numerous other decision trees. This overlap causes each individual decision tree to be less robust while permitting the whole forest of decision trees to be fairly robust. Lastly, decision trees mimic the actual decision-making process of humans [9].

3.6 Limitations

The largest limitation of decision trees is their potential overfitting of the training data. This becomes a problem when new test data is applied to the decision tree. The tree is not broad enough to encompass the new test data even though it corresponds to the training data very well. Another limitation of this model is the technique utilized in pruning the decision tree. Mingers' [10] experiment analyzed the effects of five different pruning techniques on a decision tree. The experiment found that there were significant differences between the methods. In pruning a decision tree, the resulting output may be altered depending upon which method is utilized. As for random forests, the necessity to randomize what data subset and variables are used requires some configuration and prior programming.

4 K-Nearest Neighbor

Not to be confused with K-means clustering, an unsupervised clustering technique, the K-nearest neighbor is a supervised, classification technique that is one of the simplest machine learning techniques. This model is typically used to characterize a new observation based on data that it most closely resembles or relates to. This machine learning technique can answer how we should classify data that does not have binary characteristics. There are numerous nearest neighbor algorithms to examine but specifically we will focus on the K-nearest neighbor algorithm. This model of machine learning can be developed with numerous software programs. Here we will examine how K-nearest neighbor works and what strengths and limitations are present in such models.

4.1 Use of K-Nearest Neighbor

The easiest way to visualize a nearest neighbor problem is through a two-dimensional visualization. Note however, that a nearest neighbor model can be applied to data sets with any number of variables. Figure 2 depicts a visualization of a K-nearest neighbor model with two dimension variables X_1 and X_2 . There are two classes present: orange circles and green squares. Both of these classes have already been populated through training data. The question here is should the test data, the yellow triangle near the center, be classified as a green square or an orange circle? Since this is a K-nearest neighbor model, the value of K can be manipulated. Should the value of K equal three, the yellow triangle would be classified as an orange circle since the majority of its three closest neighbors are orange circles. However, if K were to equal six, the majority of the six closest neighbors to the triangle are now green squares. Therefore, the value of K can drastically alter the results of the test data.

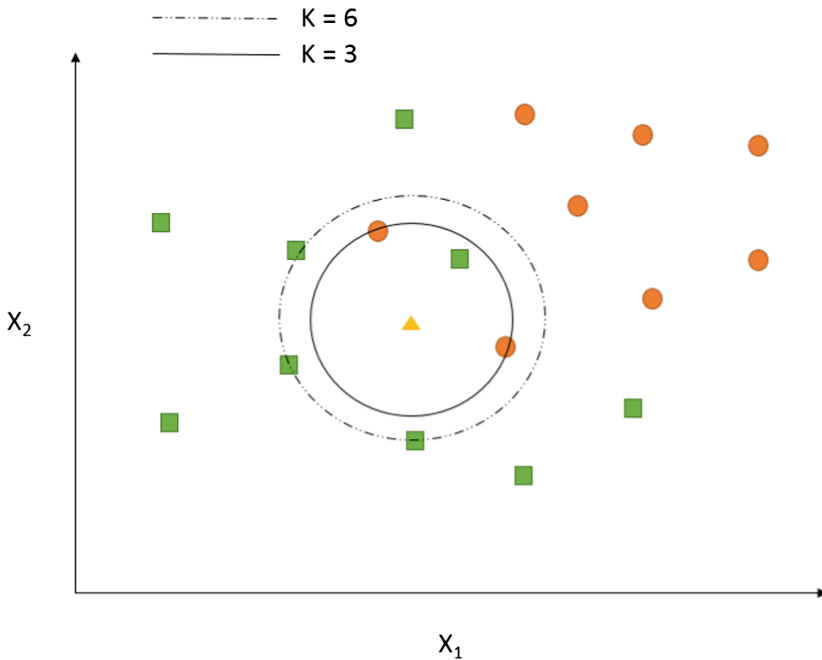


Fig. 2. Simple K-nearest neighbor Visualization

4.2 Constructing K-Nearest Neighbor (K-NN)

One of the most difficult questions to answer when conducting a K-nearest neighbor algorithm is what the value of K should be. Typically K is a smaller number due to the fact that as K increases towards the number of training points, the test point will more likely be classified as the class with the most training points. Also, a smaller K-value will result in overfitting while a larger K-value may result in oversimplification [10]. The other primary area of concern when constructing a K-nearest neighbor algorithm is determining what distance measure to use [11]. The most common distance measure is the Euclidean distance measure. However, if distance to the K-NN is large, there is an increased likelihood that an outlier would be included in the computation. Lastly, the variable ranges will need to be normalized, typically between the values of 0 and 1. This ensures that one variable does not affect the distance from the test data to the nearest neighbors more than another variable due to its range of values.

4.3 Data Required

Like decision trees, the K-nearest neighbor algorithm is non-parametric and does not rely on any assumptions on the underlying distribution of the data. This makes it a valuable algorithm especially when there is little or no prior knowledge of the data's distribution. When examining applicable data to run in a K-nearest neighbor algorithm, it is important to first look at the number of dimensions in a particular data set.

The experiment by Beyer and her colleagues [12] experiment suggests that as the number of dimensions per data set increases, the change in distance between the nearest neighbor and farthest neighbor approaches 0. This limits the data being tested to only a certain number of variables. However, data that represents more variables than computationally applicable may be able to omit some variables. The data will also need to be transformed into a set of vector inputs [12].

4.4 Strengths

Even though Nearest Neighbor is one of the simplest machine learning techniques, it is still a very strong model used to classify data. One of best qualities about Nearest Neighbor is that there are numerous enhancement techniques to classify each data set better. One example that Weinberger and colleagues [13] examined is that of a margin that takes into account other classes that are trying to invade the distance area. Another commonly used technique to weigh the distance from the test point to the K-nearest neighbors is by assigning those neighbors with a value of $(1/d)$. Also, by increasing the K-value, the model becomes more stable at the cost of potential oversimplification. Other advantages of the KNN algorithm include its ability to handle “noisy” and large training data, and the ease it which it “learns” [12].

4.5 Limitations

As depicted in Fig. 2, the greatest challenge with K-nearest neighbor is determining what the value of K should be. Too small of a value results in overfitting the data while too large of a value creates an oversimplified model. Although Euclidean distance might be the obvious choice for most K-nearest neighbor models, the distance metric may manipulate the results of different nearest neighbor models. Furthermore, K-nearest neighbor tends to be computationally expensive for many applications and result in large data sets with numerous dimensions to take a long time to compute [14]. The KNN algorithm may also be easily influenced by irrelevant attributes and tends to run slowly due to its computation complexity [12].

5 Artificial Neural Networks

Artificial Neural Network (ANN) is another machine learning tool that researchers use to solve problems. Artificial neural networks are so named because they purportedly mimic the way the human brain processes inputs and responds with an output. They comprise networks of neurons/nodes organized into layers, and synapses which are the connection in between layers [15, 17]. Then neurons are where the data are, and the synapses are the connections within the data. The result of processing with the ANN is an output or prediction, such as the “watch next” option on YouTube or the ads that appear on your web [16]. Other applications of ANNs include Amazon’s product recommendations.

5.1 How Artificial Neural Networks Work

There are many kinds of ANN architectures, one of which is the Multilayer perceptron (MLP). An MLP ANN contains at least three layers of neurons/nodes (i.e., there can be more than one hidden layer): (i) the input layer which are the predictive variables, (ii) a hidden layer (so called because they are not “visible” since they are neither the predictors/inputs nor the outputs/outcomes) which works on the data from the previous layers, and (iii) an output layer which is the outcome of the prediction. Between each adjacent layer are synapses/connections which accept the data from the multiple neurons activated from the preceding layer. The data, combined with weights, passes through an activation function, and the result determines which neurons in the subsequent layer get activated. The prediction is the result of neurons that are activated at the final output layer. During training, which can either be only in a feed-forward activation flow (uni-directional) or also include the backward propagation of errors (bi-directional), the ANN will adjust the weights and activation function of the hidden and output layers (processing layers) such that the output would most closely match the values of the outcome/target variable [17]. There are many activation functions; these include linear, step, sigmoid, linear threshold between bounds, etc. Besides the MLP, other ANN architectures include the Radial Basis Function Network, Recurrent Neural Network, the Hopfield Network, the Long/Short Term Memory Network, etc. [18].

5.2 Data Required

There is almost no limit to the type of data that can be used with ANNs, so they are used to solve a variety of problems. Data can range from being demographical information in the prediction of political affiliation, to being parts of an image for an image recognition task, to inputs from an audio file for a speech and language recognition task. However, because ANN requires numeric data, this often requires some types of data to be coded. For instance, categorical data such as “male” and “female” can be coded as “0” and “1” respectively, or the image may be coded as saturation levels in different locations on a matrix. Often, the way the data is coded impacts the quality of the prediction by the ANN. In addition to encoding, data preparation also involves standardization, which is especially necessary when nonlinear activating functions are applied. Standardization involves coding categorical or nominal data, and normalizing data.

5.3 Usability of Artificial Neural Networks

The artificial neural network can solve optimization problems, estimation problems, and cost functions just to name a few [15]. However, experts have leaned towards the idea that artificial neural networks “learn from the observed data” and act accordingly—similar to how a human brain functions [15]. Today, government agencies, small and large businesses use artificial neural networks in the most sophisticated ways possible. For example, credit card companies rely on this tool to detect any unusual activities to protect their clients and avoid fraud [16], and shipping companies use artificial neural networks to determine the fastest and most efficient route to deliver a product [16]. In

this case, shipping companies tell the artificial neural network where the product is headed and this machine learning tool takes this information, analyzes potential routes, and suggests the most cost-efficient and direct route. Other applications of ANNs are tasks in medical diagnosis, machine translation, etc.

5.4 Strengths and Limitations

Artificial neural networks' biggest strength is its versatility. Artificial neural networks can be applied to almost any scenario or problem. They implicitly address the problem of feature selection, which is one of the most challenging problems in machine learning and any prediction.

Nevertheless, just like any other machine learning tool, artificial neural networks have their limitations as well. To function at a high level, an artificial neural network must gather a large data set to operate [15]. This shortcoming can be problematic in domains where very little research has been conducted, such as extremely specific fields of study. Among the machine learning algorithms, ANNs have the weakest theoretical foundation which impedes their explanatory value since it is virtually impossible to work out the topology of the neural network – we rarely know what goes on in the hidden layers.

6 Challenges to Understanding Behavior in New Research Domains

These machine learning techniques are useful for understanding underlying patterns and can help with the prediction and classification of behaviors even when there is limited knowledge about relationships and phenomena in a domain¹. For instance, an ANN may be able to predict task-induced workload from a host of predictors that include data on demographics, personality, task characteristics, medical history, etc., but it is less useful in extracting any new constructs or measures that have no theoretical foundation. For such newer domains, where there are fewer established theories and research findings, these techniques are less able to shed light on the domain itself. To help increase knowledge and understanding in a newer domain, researchers often draw upon ideas and concepts from related domains. For instance, research in the domain of human-robot teaming has included constructs such as trust and teaming, both of which are found in social and organizational psychology, and have yielded fruitful research in human-robot teaming. This process of identifying related domains, importing constructs and research ideas from these more established and related domains is useful in spurring research in a relatively new domain and may benefit from the application of set theory, graph theory, and network analysis.

¹ In this paper, we are loosely defining “Domain” as a system comprising the increasing aggregation of units, parts, and subsystems that are interconnected and interrelated [19]. Examples of domains include the mining, nuclear plant, space missions, marine transport, power grids, manufacturing, assembly-line production domains [19].

6.1 Set Theory

Set Theory provides a way to think about elements and how they may be organized (sets). In the context of assessments in the behavioral sciences, Set Theory can be used to describe constructs (sets) in terms of their operationalization or how they are measured (elements) [20]. Since constructs in new domains tend to be less clearly defined, and their operationalization less standardized, the ability of Set Theory to deal with such ambiguity or “fuzziness” would enable even such constructs to be analyzed. Applying Set Theory may contribute to the understanding of construct/set similarity, and the degree of abstraction and generalization of constructs [20, 21].

6.2 Graph Theory

A graph is a diagram representing a system of connections or interrelations among two or more things by a number of distinctive dots, lines, bars, etc. [22]. In the assessment context, the node in the graph may be a representation of a construct set containing its measures, while a link of the graph denote the relationships among constructs. Alternatively, the nodes could be measures with the links representing the relationship among measures (see Fig. 3). Hence, Graph Theory would allow analyses of different graphs that may comprise constructs or measures, or both.

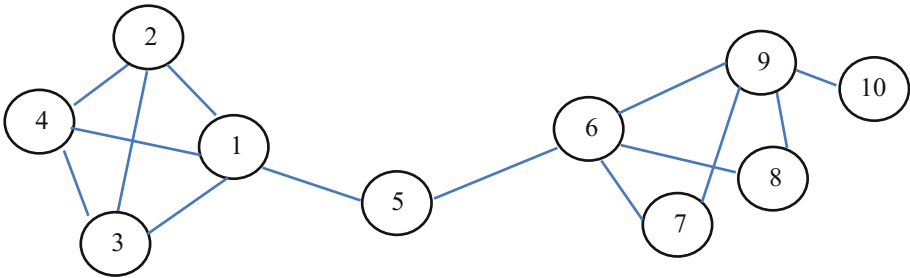


Fig. 3. A graph showing links and nodes

6.3 Network Analysis

Network science has been used in research to understand social behavior and networks. According to Lubell [23], these include:

- Identifying central individuals who can help spread ideas and behaviors (top influencers)
- Identifying disconnected individuals who need to be brought into social communities
- Identifying key social relationships that should be cultivated in order to integrate and bring together diverse communities

Network analysis and network science enable analysis of the relationships among multiple interconnected nodes and/or clusters of nodes. The nodes can be

representations of sets of various constructs and their measures, or even sets of relationships between constructs. Taking a simple example of the nodes in Fig. 3 representing constructs, with nodes 1 through 4 being constructs in a domain, and nodes 6 through 10 being constructs in another domain. Although construct/node 5 is not linked to as many construct as most of the other constructs, it enables constructs in different domains to be connected. In this case, construct/node 5 may be considered an “influential or central construct” and in network science, its *node centrality index* would be highest of all the nodes. Conversely, construct/node 10, being only linked to one other construct, may be construed as a “disconnected construct.”

7 Summary and Conclusions

This paper presented a few machine learning techniques available for data analytics. These strategies allow machines to classify, prescribe, suggest, and predict behavioral outcomes. Dependent on the situation, one technique might be more advantageous to use over another considering each technique has its own strengths and limitations. For scientists seeking to understand behaviors within and across domains, especially new domains with few established constructs and theories, methods from mathematics and network science such as set and graph theories can be beneficial. These techniques can contribute to a “bottom-up” approach, complementing the traditional “top-down” approach, to assessments and research that is more theory-driven.

Acknowledgements. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-15-2-0100. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the Army Research Laboratory of or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Jacobson, R.: 2.5 quintillion bytes of data created every day. How does CPG & Retail manage it? <https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/>
2. Marr, B.: <https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#631ddd5017b1>
3. Burford, C., Reinerman-Jones, L., Teo, G., Matthews, G., McDonnell, J., Orvis, K., Riecken, M., Hancock, P., Metevier, C.: Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness (2018)
4. Edward, J., Bagozzi, R.: On the nature and direction of relationship constructs and measurement. *Psychol. Methods* **5**, 155–174 (2000)
5. Marr, B.: Supervised V Unsupervised Machine Learning - What’s The Difference? vol. 6. <https://www.forbes.com/sites/bernardmarr/2017/03/16/supervised-v-unsupervised-machine-learning-whats-the-difference/#5ae5a61b485d>

6. Brownlee, J.: Classification and Regression Trees for Machine Learning. <http://www.machinelearningmastery.com>
7. Malakar, G.: What is Random Forest Algorithm? A Graphical Tutorial on How Random Forest Algorithm Works? <https://www.youtube.com>
8. Mitchell, T.: Decision tree learning. Machine learning, pp. 52–80. WCB/McGraw-Hill, Boston (1997)
9. Curram, S.P., Mingers, J.: Neural networks, decision tree induction and discriminant analysis: an empirical comparison. *J. Oper. Res. Soc.* **45**, 440–450 (1994)
10. Mingers, J.: An empirical comparison of pruning methods for decision tree induction. *Mach. Learn.* **4**, 227–243 (1989)
11. Wikipedia: K-nearest neighbors Algorithm. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
12. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is “nearest neighbor” meaningful? In: Beeri, C., Buneman, P. (eds.) *ICDT 1999*. LNCS, vol. 1540, pp. 217–235. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-49257-7_15
13. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* **10**, 207–244 (2009)
14. Bhatia, N.: Survey of nearest neighbor techniques. *Int. J. Comput. Sci. Inf. Secur.* **8**, 302–305 (2010)
15. Wordpress: The Shape of Data: K-nearest Neighbors. <https://shapeofdata.wordpress.com/2013/05/07/k-nearest-neighbors/>
16. Wikipedia: Artificial Neural Network. https://en.wikipedia.org/wiki/Artificial_neural_network
17. Templeton, G.: Artificial neural networks are changing the world. What are they? <https://extremetech.com/extreme/215170-artificial-neural-networks-are-changing-the-world-what-are-they>
18. Narula, G.: Machine learning algorithms for business applications- complete guide. <https://www.techemergence.com/machine-learning-algorithms-for-business-applications-complete-guide/>
19. Perrow, C.: *Normal Accidents: Living With High Risk Systems*. Basic Books, New York (1984)
20. Nelson, E.: Internal set theory: a new approach to nonstandard analysis. *Bull. Am. Math. Soc.* **83**, 1165–1198 (1977)
21. Stoll, R.R., Enderton, H.: Set Theory. <https://www.britannica.com/topic/set-theory>
22. Zweig, Katharina A.: Graph theory, social network analysis, and network science. *Network Analysis Literacy*. LNSN, pp. 23–55. Springer, Vienna (2016). https://doi.org/10.1007/978-3-7091-0741-6_2
23. Lubell, M.: Three hard questions about network science. <http://environmentalpolicy.ucdavis.edu/node/292>



A Workflow for Network Analysis-Based Structure Discovery in the Assessment Community

Grace Teo¹(✉), Lauren Reinerman-Jones¹, Mark E. Riecken²,
Joseph McDonnell³, Scott Gallant⁴, Maartje Hidalgo¹,
and Clayton W. Burford⁵

¹ Institute for Simulation and Training, University of Central Florida,
Orlando, FL, USA

{gteo, lreiner, mhidalgo}@ist.ucf.edu

² Trideum, Orlando, FL, USA

mriecken@trideum.com

³ Dynamic Animation Systems, Fairfax, VA, USA

joe.mcdonnell@d-a-s.com

⁴ Effective Applications Corporation, Orlando, FL, USA

scott@EffectiveApplications.com

⁵ Army Research Laboratory, Orlando, FL, USA

clayton.w.burford.civ@mail.mil

Abstract. When technology opens up new domains or areas of research, such as human-agent teaming, new challenges in assessments emerge. Assessments may not be as systematically conducted as new measures develop, and the research may not be as firmly grounded in theory since theories in newer domains are still being formulated. As a result, research in these domains can be fragmented. To address these, an empirically-driven network approach that is complementary to the traditional theory-driven approach is proposed. The network approach seeks to discover patterns and structure in the assessment metadata (e.g., constructs and measures) that can provide starting points and direction for future research. This paper outlines the workflow of the network approach which comprises three steps: (1) Data Preparation; (2) Data Analysis; and (3) Structure Discovery. As most of the work has been on Data Preparation, the paper will focus on the complexities and issues encountered in the first step, and include broad overviews of the subsequent steps. Anticipated use and outcomes of the network approach are also discussed.

Keywords: Network analysis · Structure discovery · Assessment Standardization · Data extraction

1 Background

1.1 Problems in Assessment: Current and Future

Research in the military and other applied fields has seen tremendous growth in the past decade [1]. With the rapid development of technology and proliferation of innovative

technological ideas, this growth in research can be partly attributed to the increase in research conducted in relatively new domains¹. One such domain is the Human-Agent Teaming (HAT) domain, where much research, especially in the military, has focused on improving human performance and informing system design [1, 3–13]. However, both at present and in the foreseeable future, there are challenges associated with research in this and other domains, particularly in the area of assessments. For example, research constructs are numerous and diverse, there is much variability in how assessments are conducted, and there are multiple contexts and timespans for assessment, among others [14].

Many of these challenges can be addressed to a large extent by (i) assessments that are more standardized and systematic, which enable stronger conclusions and findings that are more generalizable, and (ii) a strong theoretical foundation that organizes constructs to show how construct relationships are moderated by different environments and contexts. However, in these newer domains, there are fewer established theories that can provide the framework connecting the ideas and concepts underlying the research. As a result, research direction in these newer domains can tend to be overly driven by funding opportunity and expedient needs. This further exacerbates the problem since research is not conducted in a programmatic manner that facilitates building up a body of knowledge and theory development. To help address these, efforts to systematize and standardize assessments in research would be needed. In addition, a novel approach to link and cohere disparate research is proposed. The HAT domain was selected for this effort as it is a relatively new and fast-growing area of research in which many of these issues exist.

2 Overview of Network Analysis Workflow

2.1 The Network Analysis Approach

The goal of the proposed approach is to use network analysis to identify patterns and structures from the metadata in HAT research that may reveal information that can shape and improve assessments (see Table 1).

Table 1. Information that can shape and improve assessments

(a) Which constructs are most studied?
(b) Are there constructs that tend to be studied together?
(c) What measures have been used to operationalize the constructs?
(d) Are there measures more suited for certain research applications (e.g., tasks used in studies to elicit behaviors of interest) than others?

(continued)

¹ “Domain” is loosely defined as a system comprising the increasing aggregation of units, parts, and subsystems that are interconnected and interrelated [2]. Examples of domains include the mining, nuclear plant, space missions, marine transport, power grids, manufacturing, assembly-line production domains [2].

Table 1. (continued)

(e) Are there commonly-used clusters of measures?
(f) Which tasks and environments are more successful for researching the different constructs?
(g) What constructs/measures tend to be associated to which authors?
(h) Which authors collaborate within the domain? Outside the domain?
(i) Which authors tend to open up new sub-areas of research?

Such information may be used to suggest constructs that should be examined together, or measures that may be more suited for various research applications, or even potential research areas for collaboration among researchers, etc. Results of the network analysis can provide some direction for research. For instance, we may identify areas which are more and less researched, and new constructs may be suggested by patterns in clusters of measures or indicators. In contrast to the “top-down” approach of theory-driven research, this approach, which examines the links among metadata, is more of a data-driven, “bottom-up” approach.

To enable the network analysis and structure discovery, the following network approach workflow was proposed. It consists of three steps: (1) Data Preparation; (2) Data Analysis; and (3) Structure Discovery. Thus far, only the Data Preparation step has been accomplished. Hence the remainder of this paper will focus mainly on this step and brief summaries of the intended work for the other steps.

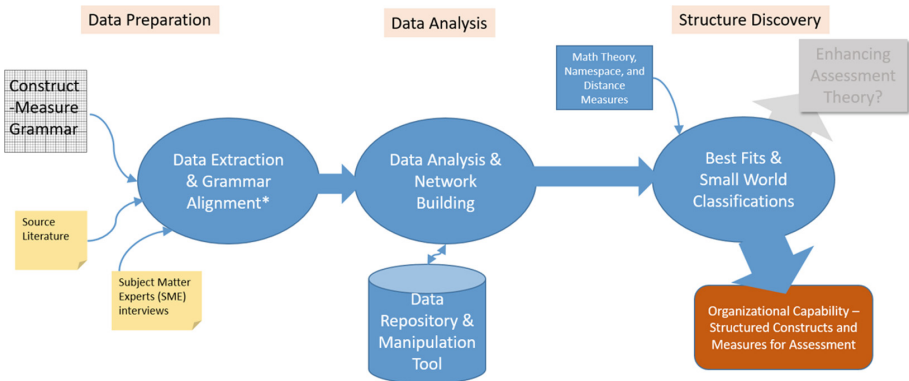


Fig. 1. Workflow of the network approach

2.2 Data Preparation

In the Data Preparation step, the goal was to extract the metadata on the assessments conducted during HAT research. To do this, we determined the boundaries of what constituted HAT research with the help of SMEs and reviewed the source literature. A construct-measure grammar was needed to help with the data extraction in this step.

2.3 Data Analysis

In the Data Analysis step, we organized the data from the Data Preparation step. We explored several database solutions that would accommodate the data type and facilitate the Structure Discovery step.

2.4 Structure Discovery

For the third step, methods and tools from network science and math (i.e., set theory, graph theory) will be utilized for the structure discovery. This is in recognition that the network approach is an exploratory attempt to derive structure from the data.

3 Data Preparation

This step involved understanding the assessment needs of researchers in the HAT domain through SME interviews, preliminary organizing of metadata to be collected with a Construct-Measure Grammar, and determining the bounds of the HAT domain and extracting the data from Source Literature.

3.1 Subject Matter Experts Interviews

We engaged researchers in the HAT domain (i.e., SMEs) to understand the current status and issues encountered in assessments so that the outcomes of the network analysis and structure discovery will be helpful in systematizing assessments and contribute to the direction of HAT research.

The SMEs were identified as researchers with notable publications on human-agent teaming, human-machine, and human-robot teaming. The semi-structured interviews were conducted with them to understand how assessments in the HAT domain have been conducted. The interviews yielded information on their areas of research interest, the variables, constructs, and measures examined in their research, the theories and models in their research, research direction, challenges and barriers faced, and funding sources, among others [14].

3.2 Construct-Measure Grammar

To address the questions in Table 1, the metadata extracted comprised information on:

- Author and co-authors
- Year of research
- Keywords (from the paper)
- Research questions
- Theories/Models cited
- Tasks and Environment used in research
- Sample characteristics

- Study design and conditions
- Variable, constructs, measures and their use in the research
- Analysis and Results

A construct-measure grammar was put in place as a means to organize the metadata in the Excel worksheet that it was saved into. The construct-measure grammar described the relationship among selected data fields. For instance, an excerpt of a row in the worksheet would read “the (i) *Independent variable*, (ii) *Transparency*, which is a (iii) *Construct* (iv) *Operationalized by* (v) *Level of Information provided by the Agent: 3 levels*, (vi) *significantly affected* the (vii) *Scores on the NASA-TLX*, which is (viii) *an operationalization of* (ix) *Workload*, a (x) *Construct* that was used as a (xi) *Dependent Variable* in the analysis.” The terms (i) to (xi) in italics are the values entered in 11 columns/fields. Data entered this way captured the relationships among the fields in the metadata.

3.3 Source Literature

The metadata of assessments in HAT research were obtained from a detailed review of HAT literature. First, the following criteria (parameters) bound the literature to be reviewed were set as:

- (a) Research within the last 5 years (i.e., 2012–2017)
- (b) Studies by DoD research laboratories (i.e., authors from the Army Research Laboratory, Navy Research Laboratory, Air Force Research Laboratory)
- (c) Empirical studies that are not meta-analyses since meta-analytic studies incorporate several studies at once, leading to double counting
- (d) **Keywords** that indicate that the study falls under the HAT domain
- (e) **Constructs** that indicate that the study falls under or is relevant to the HAT domain

Unlike the first three criteria, the last two criteria were more difficult to implement. The following sections describe the challenges, rationale, and decisions involved in using these two criteria for bounding the literature from which the metadata was extracted.

Challenge with Keywords: Multiple terms and synonyms. The working definition for HAT involved the idea of human operators working in conjunction and/or collaboration with one or more machines (e.g., robot, agent, unmanned system). However, as different researchers use different terms for similar ideas, the literature search used various synonyms for common keywords found in HAT studies. The synonyms were derived from literature and search engine suggestions.

This lack of standardization of keywords to be used in the literature search extended to the concepts in the domain. Often, there were multiple terms used for similar concepts across the studies (see Table 2). For example, in one study, “task difficulty” was analyzed as a predictor of workload, yet in another study, this has been called “task load.” Having multiple terms for the same concept would impede the network analysis and structure discovery as it would be more difficult to find common concepts and measures across studies.

Table 2. Excerpt from the thesaurus of HAT keywords and concepts

Keyword/concept	Synonyms
Human-agent teaming	<ul style="list-style-type: none"> • Human-agent teams/teaming/interaction • Human-robot teams/teaming/interaction • Human-machine teams/teaming/interaction
Unmanned systems	<ul style="list-style-type: none"> • Unmanned systems • Unmanned vehicles (UVs) • Unmanned aerial vehicles (UAVs) • Unmanned aerial systems (UASs) • Unmanned ground vehicles (UGVs) • Remotely piloted aircrafts (RPAs)
Task load	<ul style="list-style-type: none"> • Task load • Task difficulty • Task demand
Workload	<ul style="list-style-type: none"> • Workload • Mental workload • Cognitive workload • Cognitive load • Real-time workload
Stress	<ul style="list-style-type: none"> • Stress • Task-induced stress • Stress states

Decision: Creation of a Thesaurus to form Categories. Although to a large extent, the data extracted preserved the author’s original verbiage, a thesaurus of terms (i.e., keyword or concept) and the associated synonyms was needed to enable structure and patterns to be discovered from “commonalities” within the data. However, this constituted a degree of pre-processing of the data as it invariably entailed some categorization and renaming/coding of data. However, much care was taken to ensure that this pre-processing was kept to a minimum to limit any unintended “bias” that may be introduced. For instance, a term was only renamed when there was sufficient evidence (e.g., from the measures used) to indicate that it denoted the same concept as an entry in the thesaurus. A new thesaurus entry was created only for concepts that were more common and well-understood, and for which there are other known synonyms.

Challenge with Constructs: Other Constructs relevant to HAT. The first challenge with HAT constructs involved constructs from non-HAT studies. From the SME interviews and early stages in the literature review, it became apparent that some of the non-HAT studies, especially those also by HAT researchers, should be included in the literature review. These studies had examined constructs that consistently surfaced in HAT research or contributed to HAT research.

Decision: Including Other Relevant Constructs

Cyber Security

Although “cyber security” (synonym “cyber defense”) studies may not mention human-agent teaming, almost all such studies entail the human operator/cyber defender

working (teaming) with a system to identify and combat cyber threats. Given the real possibility of cyber threats and hacking of robots, agents, and systems that drive human-agent teaming, we decided to include the cyber security studies that were conducted by the researchers with published HAT research.

Driving

Another set of studies to which this rationale was applied were the driving studies published by notable HAT researchers. These studies were included as they contributed to their author's understanding of HAT since multiple HAT studies involved driving, teleoperating, or navigating with a robot.

Vigilance

Vigilance is a construct that is associated with target and threat detection, a common task in many military operations undertaken by human-agent teams. Detection tasks feature in many intelligence, surveillance, and reconnaissance missions, cyber operations, search and rescue operations, etc.

Displays and human-computer interaction/interfaces

A few HAT researchers also conducted studies on various displays and human-computer interfaces. These studies were included in the literature review because results from such studies can inform the design of the system, agent, or robot.

Challenge with Constructs: Establishment of HAT Constructs. Since a major goal for the network approach is to discover patterns and structures about constructs and measures, it was important to set the parameters of what were labeled as “constructs” in the metadata. However, this was challenging because in a relatively new domain such as the HAT domain, there can be (i) constructs which have been “borrowed” from other domains and research areas, but whose validity in the HAT domain has yet to be established, and (ii) constructs which are purportedly assessed by new researcher-developed instruments that have not been validated in as many ways. Not all constructs in the data extracted are equally well-established. Inclusion of constructs that are unestablished may affect the interpretation and “robustness” of the patterns and structures in the subsequent steps in the network approach.

Construct validation is an ongoing process and some constructs are more validated than others. Many of the issues related to assessments such as the use of different terms, and different definitions for the same construct etc., can in part be attributed to the lack of validation and establishment of some constructs. There are three broad phases in construct validation (i.e., Substantive, Structural, and External), each of which is associated with different validity evidence [15–21]. An example of *Substantive validity* evidence is literature on construct conceptualization that covers the depth and breadth of the construct [15, 21–24]. *Structural validity* can be evidenced from results of item, factor, and reliability analysis, as well as measurement invariance testing [15, 25–28]. Examples of *External validity* evidence include convergent and discriminant validity, and predictive/criterion-related validity [15, 29].

Decision: Differentiate Constructs from Variables. Although it is not possible to fully evaluate how well-established the constructs in HAT research are, we adopted the following criteria and labeled a variable as a construct if:

- It has been described as a latent variable which is associated with indicators or measures.
- It has been measured by an instrument/inventory that has especially been designed to assess it (including researcher-created inventories).
- It has been referred to as a construct by the author(s).
- It has been widely-cited as being a construct in literature.

Variables that do not meet these criteria were simply labeled as “variable.” Examples of variables include “Device type” (e.g., UAV 1 vs. UAV2), “UAS flight path type” (e.g., straight vs. figure of 8), and “Cerebral hemisphere” (left vs. right).

4 Data Analysis: Data Analysis and Network Building

Unlike the Data Preparation step, the Data Analysis and Structure Discovery steps have not been accomplished. The following sections will describe the overview of these steps.

The initial data analysis will focus on constructs and measures. It includes summarizing the data in terms of the following:

- Number and type of constructs by author
- Number and type of measures by author
- Number and type of measures by construct

The data analysis that follows may include the application of set theory, graph theory, and network analysis. The analysis can help identify the measures that are commonly used for the different constructs, and constructs that have been most-widely studied etc.

5 Data Analysis: Data Repository and Manipulation Tool

The manual data extraction resulted in an Excel file, which was converted into a .csv file. With the data at hand, the next step was to identify a database management system that (i) could accommodate the type and anticipated volume of data to be processed, (ii) would enable ease of data manipulation, and (iii) would facilitate the network analysis.

The Neo4j graph database management system [30], an open-source system written in Java, was identified as a possible system meeting these criteria. The Neo4j database schema was able to accommodate the various fields and relationships in the manually-extracted data (see Fig. 2). Neo4j has been used in knowledge graphing, social network, fraud detection, network and IT operations etc. [30].

Advantages of implementing the Neo4j as the database management system included its ability to allow easy development and operations as it required no installs, is readily configurable, and has a consistent environment. It also enables hosting of the .csv data files on a web server, facilitating file sharing. Neo4j comes with several visualization tools and allows cypher queries to be built to customize views into the data (see Fig. 3).

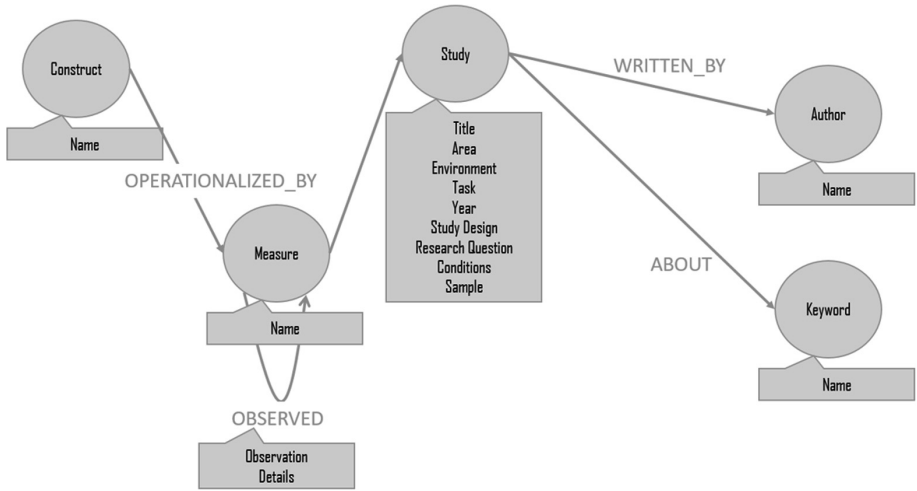


Fig. 2. Graph database schema

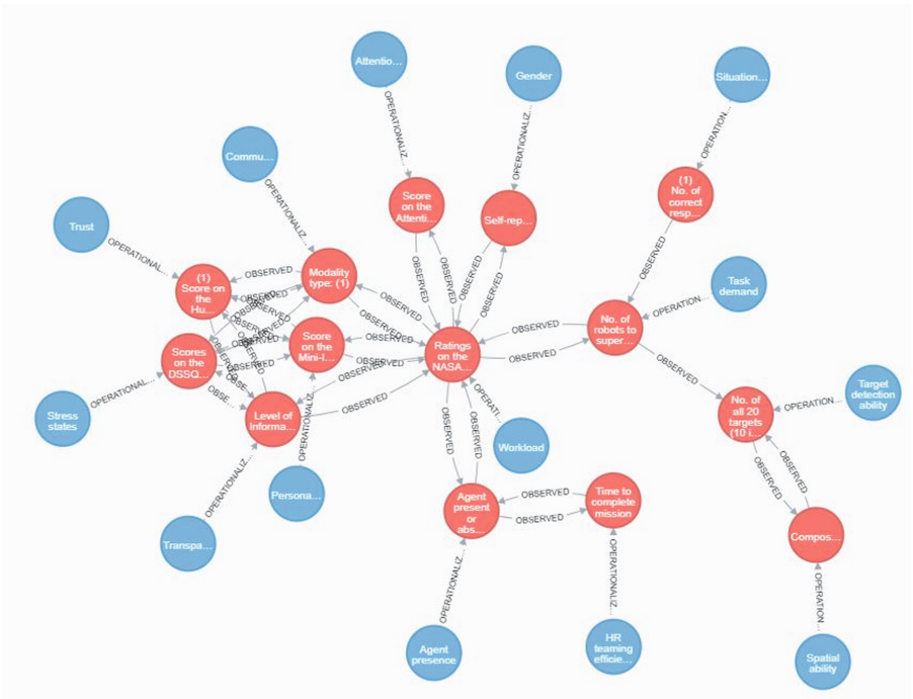


Fig. 3. Neo4j visualization of constructs and measures

6 Structure Discovery: Best Fits and Small World Classifications

6.1 Math Theory, Namespace, Distance Measures

Since there is less reliance on a theoretical framework, data analysis is expected to be iterative and the results subject to close scrutiny by researchers. This “bottom-up” approach seeks to discover a posteriori (e.g., bottom-up) data schemes. The math will allow multiple schemes to be generated, although it is important that not many assumptions be made about the underlying order or structure that may bias structure discovery. The team intends to attempt multiple schemes and look for “best fit,” based on inputs from researchers.

If the network approach results in “discoverable structure” in the data, we would then seek to:

- Find not just one set of relationships and patterns in the data, but many, or all that are there
- Find relationships that will allow us to suggest where new data fits in
- Find relationships that allow us to better categorize research
- Find relationships that may expose or confirm real-world cause and effect in assessment phenomena of interest, especially constructs and their influencers
- Find relationships that suggest potential areas for collaboration among researchers

There are currently two methods to be explored: (i) graph theory [31] and network analysis [32], and (ii) set theory [33, 34]. These methods were mainly selected as they are “scale-free” and do not impose assumptions about the data. Table 3 outlines the similarities and differences between these methods.

Table 3. Two possible methods for Structure Discovery

Graph theory and Network analysis	Set theory
<ul style="list-style-type: none"> • Identify authors, concepts, constructs, etc., as nodes • Look for relationships among and across them • Results must be interpreted • Certain properties of networks are also “scale-free” (e.g., small worlds) thereby providing a degree of continuity for this research framework as we incorporate more and more data 	<ul style="list-style-type: none"> • Describe constructs via set of operationalizations • Consider abstraction and generalization • Explore various set similarity measures • Can be complemented by other related mathematical techniques such as the fuzzy set theory and set similarity
<ul style="list-style-type: none"> • Possesses “neutral” organizing principles that enables an objective “lens” whereby to view the state of the assessments • By not forcing a rigid theoretical structure, method allows for continual adjustments and “newcomers” • Method can result in structures that further point to underlying concepts that in turn may lead to increased theoretical systemization 	

7 Conclusion

The network approach was proposed to address the specific challenges with current and future assessments in newer domains. These are the lack of standardization and systematization of assessments and the presence of fragmented research that resulted from the absence of strong theories to cohere research. The approach is an empirical, “bottom-up” approach that seeks to discover structures and patterns in the metadata that may provide insight into underlying relationships among constructs and concepts that can suggest related areas of study. Anticipated outcomes of the network approach include information about which measures have been most associated with which constructs, which researchers have similar research interests and may be potential collaborators, and which constructs could be examined together.

Acknowledgements. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-15-2-0100. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the Army Research Laboratory of or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Tsarouchi, P., Makris, S., Chryssolouris, G.: Human–robot interaction review and challenges on task planning and programming. *Int. J. Comput. Integr. Manuf.* **29**, 916–931 (2016)
2. Perrow, C.: *Normal Accidents: Living with High Risk Systems*. Basic Books, New York (1984)
3. Barnes, M.J., Chen, J.Y.C., Jentsch, F., Redden, E.S.: Designing effective soldier-robot teams in complex environments: training, interfaces, and individual differences. In: Harris, D. (ed.) *EPCE 2011. LNCS (LNAI)*, vol. 6781, pp. 484–493. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21741-8_51
4. Chen, J.Y.C., Haas, E.C., Barnes, M.J.: Human performance issues and user interface design for teleoperated robots. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **37**, 1231–1245 (2007)
5. Endsley, M.R., Jones, W.M.: A model of inter- and intra-team situation awareness: Implications for design, training and measurement. In: *New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments*, pp. 46–67. Human Factors and Ergonomics Society, Santa Monica (2001)
6. Goodrich, M.A., Olsen, D.R.: Seven principles of efficient human robot interaction. In: *2003 IEEE International Conference on Presented at the Systems, Man and Cybernetics* (2003)
7. Groom, V., Nass, C.: Can robots be teammates? benchmarks in human–robot teams. *Interact. Stud.* **8**, 483–500 (2007)
8. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. *Hum. Factors* **46**, 50–80 (2004)
9. Lyons, J.B., Sadler, G.G., Koltai, K., Battiste, H., Ho, N.T., Hoffmann, L.C., Smith, D., Johnson, W., Shively, R.: Shaping trust through transparent design: theoretical and experimental guidelines. In: *Savage-Knepshild, P., Chen, J. (eds.) Advances in Human Factors in Robots and Unmanned Systems*, pp. 127–136. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-41959-6_11

10. Millot, P., Pacaux-Lemoine, M.-P.: A common work space for a mutual enrichment of human-machine cooperation and team-situation awareness. *IFAC Proc.* **46**, 387–394 (2013)
11. Sibley, C., Coyne, J., Morrison, J.: Research considerations for managing future unmanned systems. Naval Research Laboratory, Washington, United States (2015)
12. Gu, W., Mittu, R., Marble, J., Taylor, G., Sibley, C., Coyne, J., Lawless, W.F.: Towards modeling the behavior of autonomous systems and humans for trusted operations. In: Presented at the 2014 AAAI Spring Symposium Series (2014)
13. Yanco, H.A., Drury, J.L., Scholtz, J.: Beyond usability evaluation: analysis of human-robot interaction at a major robotics competition. *Hum.-Comput. Interact.* **19**, 117–149 (2004)
14. Burford, C.; Reinerman-Jones, L., Teo, G., Matthews, G., McDonnell, J., Orvis, K., Riecken, M., Hancock, P., Metevier, C.: Unified Multimodal Measurement for Performance Indication Research, Evaluation, and Effectiveness (2018)
15. Flake, J.K., Pek, J., Hehman, E.: Construct validation in social and personality research: current practice and recommendations. *Soc. Psychol. Pers. Sci.* **8**, 370–378 (2017)
16. Benson, J.: Developing a strong program of construct validation: a test anxiety example. *Educ. Meas. Issues Pract.* **17**, 10–17 (1998)
17. Clark, L.A., Watson, D.: Constructing validity: basic issues in objective scale development. *Psychol. Assess.* **7**, 309–319 (1995)
18. Crocker, L.M., Algina, J.: Introduction to Classical and Modern Test Theory. Wadsworth Publishing Company, Belmont (2006)
19. Loevinger, J.: Objective tests as instruments of psychological theory. *Psychol. Rep.* **3**, 635–694 (1957)
20. Raykov, T., Marcoulides, G.A.: Introduction to Psychometric Theory. Routledge, New York (2011)
21. Gehlbach, H., Brinkworth, M.E.: Measure twice, cut down error: a process for enhancing the validity of survey scales. *Rev. Gen. Psychol.* **15**, 380–387 (2011)
22. Dawis, R.V.: Scale construction. *J. Couns. Psychol.* **34**, 481–489 (1987)
23. Willis, G.B.: Cognitive Interviewing: A Tool for Improving Questionnaire Design. Sage Publications, Thousand Oaks (2004)
24. Sireci, S.G.: The construct of content validity. *Social indicators research*, pp. 83–117. Kluwer Academic Publishers, Netherlands (1998)
25. McDonald, R.: Test homogeneity, reliability, and generalizability. In: *Test Theory: A Unified Approach*, pp. 76–120. Lawrence Erlbaum Associates, Mahwah (1999)
26. McCrae, R.R., Kurtz, J.E., Yamagata, S., Terracciano, A.: Internal consistency, retest reliability, and their implications for personality scale validity. *Pers. Soc. Psychol. Rev.* **15**, 28–50 (2011)
27. Chmielewski, M., Watson, D.: What is being assessed and why it matters: The impact of transient error on trait research. *J. Pers. Soc. Psychol.* **97**, 186–202 (2009)
28. Millsap, R.E.: *Statistical Approaches to Measurement Invariance*. Routledge, Florence (2012)
29. Campbell, D.T., Fiske, D.W.: Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **56**, 81–105 (1959)
30. Neo4j. <https://neo4j.com/>
31. West, D.B.: *Introduction to Graph Theory*. Prentice hall, Upper Saddle River (2001)
32. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. *Science* **323**, 892–895 (2009)
33. Nelson, E.: Internal set theory: a new approach to nonstandard analysis. *Bull. Am. Math. Soc.* **83**, 1165–1198 (1977)
34. Moore, R., Lodwick, W.: Interval analysis and fuzzy set theory. *Fuzzy Sets Syst.* **135**, 5–9 (2003)

Augmented Cognition in Virtual and Mixed Reality



Immersion Versus Embodiment: Embodied Cognition for Immersive Analytics in Mixed Reality Environments

Denis Gračanin^(✉)

Department of Computer Science, Virginia Tech,
Blacksburg, VA 24060, USA
gracanin@vt.edu

Abstract. Visualization techniques are used to analyze and understand data within the context of the underlying conceptual and physical model. We collect and generate data at an increasingly fast rate, but visual analysis capabilities are lagging. The challenges of visual data analysis and exploration are associated with very large data sets, increased dimensionality, and the consideration of data semantics, including features, focus and context. Typically, visual analysis is done using Coordinate Multiple Views (CMV) tools that support linking and brushing (selection) of data in multiple synchronized views. The recent advances in MR technologies provide a great opportunity to support deployment and use of MR applications for visualization and visual analytics. Direct mapping of CMV tools to an MR environment arguably creates more problems than it solves. Embodied interactions and embodied user interfaces lead towards invisible user interfaces and move the visualization and analysis from a computer screen to physical space and place. It is necessary to explore various interaction and visualization modalities in MR environments to identify best practices to leverage embodied cognition and interactions. Such explorations can benefit from a framework and an evaluation testbed for embodied interactive immersive analysis. The framework provides services for data access, data visualization (views), both traditional two-dimensional view and a three-dimensional equivalent, views assembly, gestures, and interaction devices. An Internet of Things based Smart Built Environment example is used to illustrate the proposed approach.

Keywords: Mixed reality · Immersive analytics

1 Introduction

Visualization techniques help us analyze and understand data within the context of the underlying conceptual and physical model. However, there is still a cognitive gap that needs to be closed to facilitate better interactive visual analysis. More sophisticated approaches use immersive virtual environments [34], game

engines [5], and augmented/mixed reality [11]. Those approaches focus mostly on the visual aspects without taking into account the importance of the user's interactions with the physical world.

Typically, visual analysis is done using a coordinate multiple views (CMV) tool that supports linking and brushing (selection) of data in a variety of view types [35]. The synergy of multiple views, each showing the brushed data from different perspective, facilitates insight into a complex data set.

Mixed reality (MR) [7] incorporates digital artifacts in the surrounding physical world creating an MR environment that implicitly leverages embodiment and affordances of the physical world while providing the visualization capabilities of the digital world. If those digital artifacts, or virtual objects, are data representatives, we can use their placements in the physical world to reduce the cognitive gap and enhance the user's ability to analyze the data. In other words, we can bring the CMV paradigm into an MR world.

MR technologies allow for providing users with an environment that blends the physical surroundings with virtual objects. The recent advances in MR technologies provide a great opportunity to support the deployment and use of MR applications for visualization, simulation, training, and education, in single-user and collaborative settings. Users can interact with virtual objects that can help them be more engaged and acquire more information compared to the more traditional approaches.

The most obvious approach is to directly map the CMV tool views to "floating" digital screens in an MR environment and implement brushing using a simple point and click interaction paradigm. While such an approach is relatively easy to implement, it arguably creates more problems than it solves. The problems include technical (MR device resolution, field of view, spatial mapping, occlusion), usability (two-dimensional views in a three-dimensional space) and cognitive (switching between the views, grasping the overall view of data). Immersion in an MR environment does not necessarily benefits from embodiment.

Therefore, the challenge is to explore various interaction and visualization modalities in an immersive MR environment to identify what are the best practices to leverage embodied cognition and interactions to support immersive visual analytics. In order to address and explore that challenge, we developed a framework and its initial reference implementation as an evaluation testbed for embodied interactive immersive analysis. The goal is to take advantage of affordances and embodied cognition in an MR environment.

The framework provides services for data access, data visualization (views), both traditional two-dimensional view and a three-dimensional equivalent, views assembly, gestures, and interaction devices. The implementation is based on a Microsoft HoloLens device combined with eye-tracking, tracking, and biometric wrist band devices to provide for accurate monitoring of user actions, eye gaze and physiological signals.

We use an Internet of Things (IoT) [36] based Smart Built Environment (SBE) physical system to illustrate the framework and to evaluate the testbed. The historical and current sensor data are stored in a data repository and

accessed based on the context and user interest. Visual analysis can be conducted in the original SBE or in any other space. A pilot user study is under way to test the functionality and services provided by the framework as well as to refine and extend the provided services. The survey data, task performance, and the physiological data will be used to evaluate the cognitive gap and its effects based on the level of embodiment.

2 Related Work

One of the challenges of visual analytics is how to analyze a huge amount of heterogeneous data [37]. Interactive visual analytics is often limited by our inability to fully grasp the data presented on a computer screen due to the cognitive overload. User interaction and ability to quickly search, filter, visualize, and analyze data is essential. Supporting a high-level information processing requires that a user can filter, visualize, and navigate data [8].

Interactive brushing or dynamic querying enables iterative on-the-fly formulation and refinements of data analysis tasks based on the visual feedback provided by multiple views [15]. CMVs are frequently used for interactive visual analysis because they provide multiple simultaneous perspectives or view of the data [35]. There are many available CMV tools tailored for different application domains [29].

Our cognitive processes depend on how our body interacts with the physical world (affordances) and how we off-load cognitive work onto our physical surrounding (embodied cognition). Embodiment cognition [39] leverages the notion of affordances, potential interactions with the environment, to support cognitive processes. Embodied interactions [12] and embodied user interfaces [16] lead towards invisible user interfaces moving the computation and analysis from computer screen to a physical environment [13,38]. Embodied interactions demonstrate the importance of the body's interactions with the physical world.

Three-dimensional (3D) Virtual Reality (VR) environments allow users to explore virtual worlds without actually "being there." VRs have been used in a variety of applications including education, training, architectural walkthroughs, scientific visualization, art, and entertainment. Although VR applications can be used for many different purposes, there are some fundamental interaction tasks used in VR applications.

VR applications can be used to map or represent real-world sensors and embedded devices to virtual devices and objects. VR-based user interface and visualization are used to evaluate IoT computing applications and to interactively test various IoT configurations [14,30,33].

Experience from 3D user interfaces and interactions with VR environments can be a starting point for interaction with the real world and architectural artifacts [21,26]. By providing an MR environment that uses the surrounding architectural space as a context for visualization and is connected to controls, sensors, and actuators in the architectural space, we can use human-computer interaction based approaches and interaction techniques to support MR interactive visualization.

MR environments are also well suited for tangible visual analysis because the physical objects (props) can be used to interact with the physical world and with the virtual objects, thus leveraging embodiment and context awareness [19, 24, 25]. New MR technologies (e.g., Microsoft HoloLens device) can augment the user experience and provide affordances for brushing by allowing the user to define the context using semantic representatives (virtual/tangible and 2D/3D). We can use the concept of tangible brushing for exploratory analysis in a MR environment using tangible (semantic) representatives for data dimensions [20]. Manipulation of representatives selects a subset of data providing the users with immediate feedback and a powerful support for data comprehension.

MR environments can be used in a single-user and in a collaborative, group settings. MR systems are well suited for collaborative and distributed work because users interact in face-to-face mode with the real world as well as the virtual objects even if they are not co-located. These results are reflected in collaborative and standalone MR applications. The applications range from robotics [17] and manufacturing [18] to gaming [23] and visualization [40].

However, a collaboration of multiple users who are not necessarily co-located introduces additional challenges. Billinghurst and Kato explored the notion of functional and cognitive seams in collaborative MR systems [6] and reviewed MR techniques for developing collaborative interfaces.

Contextual information about the physical surrounding can help improve MR experience. The IoT paradigm provides integration between the physical and the digital worlds and allows for new applications that can benefit from connecting everyday objects to the internet. In 2008, the US National Intelligence Council (NIC) included IoT in a list of six disruptive civil technologies with potential impact on US interests.

The IoT architecture and the corresponding implementation [10] differ from the traditional network architecture. A large number of devices are connected, most of them with limited computing and networking capabilities. The IoT devices are deployed in various contexts [1], including wearable devices, house appliances/sensors, embedded devices/smartphones, SBEs, and environmental sensors. They can span large urban areas to support “smart cities” [27]. With an intelligent infrastructure core, large number of sensors, and mobile, ubiquitous access, IoT provides many opportunities for innovation. Examples include MR human-human, human-device and device-device collaborations, personalized healthcare/medicine, intelligent transportation with autonomous and semi-autonomous vehicles, and SBEs for various living and work settings.

3 Visual Analytics, Immersion, and Embodiment

Thomas and Cook [37] define visual analytics as “the science of analytical reasoning facilitated by interactive visual interfaces.” It is a wide-ranging field of science that involves visualization and interaction methods combined with analytical reasoning, data representation and transformation as well as production and presentation of the results.

We are able to collect and generate data at an increasingly fast rate, but capability of analyzing the collected data lags behind. We focus on how analysts gain insight into data, find expected and unexpected features, and make decisions using visual tools. The analysts' main goal is always to explore phenomena and test hypotheses or to discover unexpected results that question established assumptions or the validity of the data acquisition process. That can lead to the generation of new hypotheses.

The challenges of data analysis and exploration are associated with very large data sets, increased dimensionality and the consideration of data semantics, including features, focus, and context. Therefore, a visualization tool should be designed in close collaboration with potential users.

When dealing with a large amount of information, we first find and understand individual pieces of information and then develop a combined understanding of a data set containing many heterogeneous pieces of information. Such activities provide insight formation (comprehension, making sense, storytelling, or interpretation). The goal is to get new insights through the novel combination, organization, or structuring of known information. This insight formation is crucial for individual users as well as for groups of users.

It is important to explore and understand how analysts cope with complex models in various tasks and how can they benefit from collaboration. However, not all analysts should be visualization experts. Rather, they should be provided with semantic interfaces that can adjust to their needs and common knowledge.

Insight formation is affected by the organization and coordination of views and user interactions with those views. Therefore, view management and input modalities play an important role in both traditional visualization and VR/MR visualization. Most of the visual analytical systems use either a traditional desktop or large scale displays with direct touch.

However, in VR/MR view management involves maintaining visual constraints on the 'data' objects, locating related objects near each other, or preventing occlusion [4]. Affordances of different input modalities (touch, speech, proxemics, gestures, gaze, and wearable) can be described in terms of direct interactions (no mediator) between a human body and 'data' objects [3]. For example, spatial immersion introduces challenges such as depth perception, data localization and object relations [28].

Immersive analytics could provide the ultimate user interface for insight formation within integrated immersive data worlds. For that, it is essential to provide a mix of creative user interactions, support for collaboration, and insightful learning algorithms [22].

There is a lack of methods and practical guidelines for the development of embodied user interfaces, especially in the context of MR environments. Much has been written about embodiment theory, and there are examples of effective embodied interfaces, but how can we approach the design and implementation of a user interface or an interaction technique that would enable users to make use of their powerful embodied resources in an MR environment? The ability

to combine immersion and embodiment in MR environments is critical to the success of the MR-based immersive analytics.

We can make embodiment more concrete by using the theory of affordances [32]. We can investigate the affordances provided by various user interface designs, and to evaluate which affordances lead to a higher degree of “embodied behavior,” a behavior that indicates the user is taking advantage of embodied resources to gain better insight, especially in a collaborative setup [9].

In order to investigate impact of embodiment and immersion on visual analytic in MR environments, we need to developed the supporting infrastructure. This infrastructure include a testbed to develop and deploy applications and a framework that provides data collection, fusion and evaluation services.

Since the future visual analytics infrastructure will be distributed and collaborative [31], we need to develop collaborative frameworks that will bring together geographically distributed and co-located analysts. Only through gradual refinement and improvements that takes advantage of recent immersive technologies, we can provide new immersive analytics, control, interaction, and visualization modalities.

However, most MR-enabled collaborative work done over the network has seams, namely spatial, functional, and cognitive seams [6]. Spatial seams are a consequence of geographically separated and technologically asymmetric spaces of individual users. Functional seams steam from different functional workspaces, forcing the user to change operation. Cognitive seams result from differences between existing and new work practices. Therefore, it is necessary to address those seams and enhance the collaborative experience in order for the MR systems to go beyond what current systems could do.

4 Immersive Analytics Frameworks and Services

The need for new immersive analytics, control, interaction, and visualization modalities is only one side of the coin. The data about the analysts, especially cognitive state and stress level, can direct how the data objects are presented and interacted with. An instrumented physical space enriched with a variety of sensors can provide a lot of data about the user and the surrounding physical environment. The environmental and physiological data can be combined with self-reported data to create a contextualized data about the user. This contextualized data then informs the stimuli that is provided to the user (Fig. 1).

The capabilities are provided through basic environmental, physiological and cognitive data retrieval service. The derived composite service implement algorithms and techniques to provide activity, stress and affect related information. The services help us explore how to design human-space interaction, how to identify the modalities of interactions, and how to inform the overall design of space when it is superimposed with technology.

The challenge is how to effectively collect, analyze and fuse data while managing heterogeneous and geographically distributed users. Creating a MR collaborative system that takes advantage of fused data and orchestrated interactions

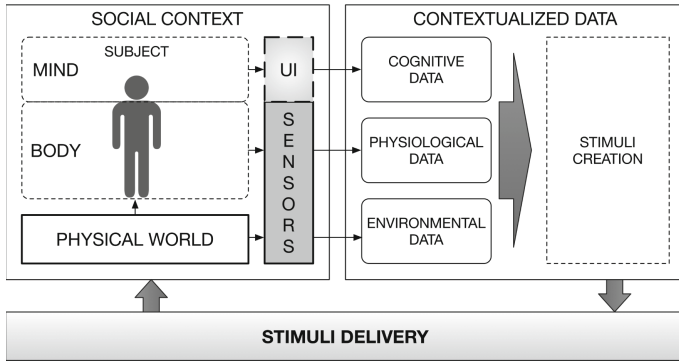


Fig. 1. Contextual physical environment: an instrumented physical space.

can improve quality of experience and interactions, and reduce spatial, functional and cognitive seams for more effective and efficient human networks.

The present extent of context-aware applications leaves a lot to be desired. The problem is that context awareness shown by human beings is based on a radically different paradigm as compared to the one used by the computational infrastructure.

Figure 2 shows that a typical physical space (a classroom) can be quickly transformed in a hospital room, bathroom or a model of an SBE. Due to the immersion and embodiment of MR environments, the context changes dramatically due to a closed loop from the user to stimuli and then back to the user.



Fig. 2. Transforming an instrumented physical space into a MR environment **Left:** Physical environment, a classroom. **Right:** MR environment, a hospital room.

Creative user interactions [22] must take advantage of the human body as an interaction devices. However, the presence of tangible, reconfigurable smart objects, provide tremendous opportunities to address the needs of the immersive analytics process in MR environments.

Unlike traditional graphical user interfaces (WIMP paradigm), tangible user interfaces are not well integrated. There is a cognitive and contextual gap due to the space separation between the user input and display. Various modalities of tangible inputs (e.g., AR markers, smart objects) must be integrated with a display within the user interface space. Visually overlaying tangible input with visual analysis display within an MR environment must address issues such as occlusion, limited embodiment and limited user interface space [20]. Figure 3 shows how tangible interactions for MR analytics can be created by combining a traditional CMV tool with tangible markers to provide a tangible input space.

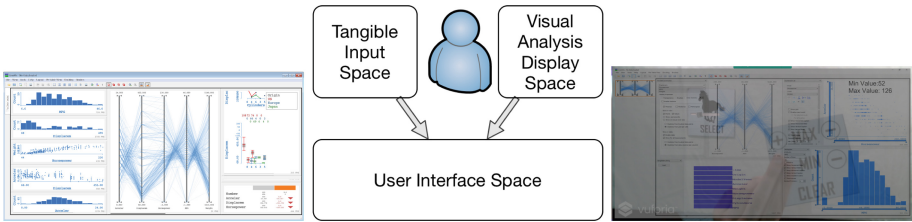


Fig. 3. Tangible interactions for MR analytics. **Left:** A traditional CMV tool. **Center:** The user interface space is a combination of a tangible input space and a visual analysis display space. **Right:** The implementation of the user interface space by integrating the output of the traditional CMV tool within an MR environment.

We have to move beyond the 2D displays and 2D representatives to leverage benefits of embodied interactions. MR devices with integrated cameras and environment mapping capabilities can provide support for flexible 3D tangible and 3D virtual representatives. In that context, a smart tangible object can serve as prop for a number of interactions. Figure 4 shows such a smart, reconfigurable tangible object representing an adjustable money bill. Using smart tangible objects reduces the need for artificial, learned gestures and provides affordances for embodied interactions. The smart, tangible object becomes an intermediary for interactions between the user and the avatar or between geographically distributed users who see each other as avatars.

Using the physiological bio-sensing devices (wearable sensors for galvanic skin-response, heart-beat, blood volume pulse, etc.), arguably captures subconscious processes in the human body. Therefore, such approach is unlikely to interfere with the user’s activities due to social or other considerations. That, in turn, allows exploration of social and technology interactions in information-rich physical spaces and places that characterize MR environments.

Figure 5 shows a local user (left), a remote user (center) and an avatar controlled by the application (right). The environmental data (tracking) is used to position the remote use in the local user’s space. The avatar representation of the remote user mirrors the remote user’s body movement. A simple visualization of the remote user’s biometric data, blood volume pulse (BVP), galvanic skin

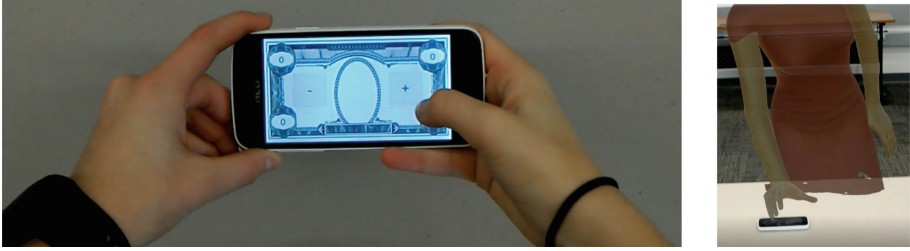


Fig. 4. Left: A view of a smart, tangible object representing an adjustable money bill. The user can set the amount using the ‘-’ and ‘+’ icons. Right: The avatar reaching towards the smart object to indicate setting the return value.



Fig. 5. Remote collaboration. **Left:** A local user in a distributed MR environment. **Center:** A remote user is tracked and represent by an avatar that replicates, in real-time, the remote user’s body movements and displays biometric data. **Right:** An avatar that is controlled by the application to interact with the user.

resistance (GSR), and temperature (TMP) provides an additional context and possible indication of the remote user’s affect.

4.1 Implementation

The described characteristics of immersive analytics frameworks and related services demonstrate the current stage in the development of a framework and an evaluation testbed for embodied, MR-based immersive analysis. The framework provides services for data access, data visualization (views), both traditional two-dimensional view and a three-dimensional equivalent, views assembly, gestures, and interaction devices. The provided services include, among others:

- Collection and visualization of body tracking data using an avatar in an MR environment.
- Visualization of sensor data (environmental and biometric) within the MR environment.
- Data exchange among multiple sites that are aligned in a common MR environment.
- Support collaborative interactions and manipulation in the MR environment.

The collaboration and access to the remote tracking and sensor data is implemented using MQTT, a popular lightweight M2M communication protocol for exchange of telemetry data [2]. The data is stored in a data repository (OSIsoft PI system). The collected longitudinal data can be accessed, viewed and analyzed using the framework services.

4.2 Example: Smart Built Environments

A smart environment is a physical space enriched with smart objects that work continuously to make residents lives more comfortable SBEs incorporate sensors and actuators into the built space to provide new functionalities or enhance the current ones.

User tasks in built environments usually involve interaction with physical objects that respond to user actions with some kind of feedback. In addition, the user can view and analyze SBE data. Services allow users to both issue commands and receive feedback. The services exist on multiple scales ranging from the design of door knobs to the placement of house modules.

SBE augments traditional home by adapting new technology into the existing patterns of use to provide a rich computational and communication infrastructure. This infrastructure includes smart things, devices and sensors that can observe the physical environment and interact with the inhabitants in novel ways. However, the physical and social structures within a built environment are subject to continuous change that creates the need for reconfigurable spaces and places in SBEs.

An ongoing FutureHAUS project focuses on the design and construction of a smart, modular house. The framework was used to develop and instrument a supporting hybrid testbed for the visualization and analysis of the smart house's modules. Among others, testbed was used to implement and explore new modalities in using lighting in built environments. The conducted user studies investigated the impact of visual stimuli on cognitive states and emotions.

Figure 6 left shows visual representation of an SBE (a kitchen module) in a large room to provide for a realistic walkthrough (in actual size). Figure 6 right shows positioning an avatar in a public space (library).



Fig. 6. Positioning virtual object in physical space: **Left:** A model of a kitchen module (actual size) placed in the Cube facility space. **Right:** An avatar positioned in a library.



Fig. 7. Kitchen module: **Left:** The constructed kitchen module, a part of a modular smart house (SBE). **Right:** The VR representation of the kitchen module.

Figure 7 shows the completed kitchen model (left) and its VR equivalent (right).

5 Conclusion

Immersive analytics has tremendous potential in addressing visual analytics challenges. However, that will require new immersive analytics, control, interaction, and visualization modalities. The challenge is to identify what are the best practices to leverage embodied cognition and interactions to support immersive visual analytics.

Immersion alone is not sufficient and must be complemented by embodiment. Fully immersive VR environments isolate the user from the physical world and drastically reduce affordances for embodied interactions. MR environments provide sufficient immersion while maintaining the affordances for embodied interactions. The presented framework and the corresponding testbed provide services for embodied interactive immersive analysis. The goal is to take advantage of affordances and embodied cognition in an MR environment.

Using the framework based tools in a real-world context, where computer generated stimuli are blended with the real-world stimuli through the use MR technologies, will provide necessary affordances in support of higher-fidelity, embodied interactions that will, in turn, result in more effective visualization and analysis.

References

1. Anderson, J., Rainie, L., Duggan, M.: The internet of things will thrive by 2025. Technical report, Pew Research Center, Washington, D.C., May 2014
2. Arlitt, M., Marwah, M., Bellala, G., Shah, A., Healey, J., Vandiver, B.: MQTT version 3.1.1 plus errata 01. Standard, OASIS, 10 December 2015

3. Badam, S.K., Srinivasan, A., Elmqvist, N., Stasko, J.: Affordances of input modalities for visual data exploration in immersive environments. In: Proceedings of the Workshop on Immersive Analytics: Exploring Future Interaction and Visualization Technologies for Data Analytics (#Immersive 2017) – IEEE VIS (2017)
4. Bell, B., Steven Feiner, S., Höllerer, T.: View management for virtual and augmented reality. In: Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology, pp. 101–110. ACM, New York (2001)
5. Bijl, J.L., Boer, C.A.: Advanced 3D visualization for simulation using game technology. In: Proceedings of the Winter Simulation Conference, pp. 2815–2826 (2011)
6. Billinghamurst, M., Kato, H.: Collaborative mixed reality. In: Ohta, Y., Tamura, H. (eds.) International Symposium on Mixed Reality (ISMR 1999), pp. 261–284. Springer, Heidelberg (1999)
7. Bimber, O., Raskar, R.: Spatial Augmented Reality: Merging Real and Virtual Worlds. A K Peters, Wellesley (2005)
8. Card, S.K., Mackinlay, J., Shneiderman, B. (eds.): Readings in Information Visualization: Using Vision to Think. Interactive Technologies, Morgan Kaufmann, San Francisco (1999)
9. Clark, A.: Embodied, situated, and distributed cognition. In: Bechtel, W., Graham, G., Balota, D.A. (eds.) A Companion to Cognitive Science, pp. 506–517. Blackwell Publishing Ltd., Oxford (2017)
10. daCosta, F.: Rethinking the Internet of Things: A Scalable Approach to Connecting Everything. Apress L. P., Berkeley (2013)
11. Dong, S., Kamat, V.R.: Collaborative visualization of simulated processes using tabletop fiducial augmented reality. In: Proceedings of the Winter Simulation Conference, pp. 828–837 (2011)
12. Dourish, P.: Where the Action Is: The Foundations of Embodied Interaction. The MIT Press, Cambridge (2001)
13. Dourish, P.: Re-space-ing place: “place” and “space” ten years on. In: Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW 2006), pp. 299–308. ACM, New York, 4–8 November 2006
14. Eastman, C., Teicholz, P., Sacks, R., Liston, K.: BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Architects, Engineers, Contractors, and Fabricators. Wiley, Hoboken (2008)
15. Fishkin, K., Stone, M.C.: Enhanced dynamic queries via movable filters. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 415–420. ACM Press/Addison-Wesley Publishing Co., New York (1995)
16. Fishkin, K.P., Gujar, A., Harrison, B.L., Moran, T.P., Want, R.: Embodied user interfaces for really direct manipulation. *Commun. ACM* **43**(9), 74–80 (2000)
17. Frank, J.A., Krishnamoorthy, S.P., Kapila, V.: Toward mobile mixed-reality interaction with multi-robot systems. *IEEE Robot. Autom. Lett.* **2**(4), 1901–1908 (2017)
18. Gonzalez-Franco, M., Pizarro, R., Cermeron, J., Li, K., Thorn, J., Hutabarat, W., Tiwari, A., Bermell-Garcia, P.: Immersive mixed reality for manufacturing training. *Front. Robot. AI* **4**, 3 (2017)
19. Gračanin, D., Eck II, T., Silverman, R., Heivilin, A., Meacham, S.: An approach to embodied interactive visual steering: bridging simulated and real worlds. In: Proceedings of the 2014 Winter Simulation Conference (WSC), pp. 4073–4074, December 2014

20. Gračanin, D., Tasooji, R., Handosa, M., Matković, K., Waldner, M.: Tangible visual analysis: brushing in a mixed-reality environment. In: Puig, A., Isenberg, T. (eds.) *Proceedings of the 19th EG/VGTC Conference on Visualization (EuroVis 2017) – Posters*. The Eurographics Association, 12–16 June 2017
21. Gračanin, D., Zhang, X.: CaffeNeve: a podcasting-capable framework for creating flexible and extensible 3D applications. *ACM Comput. Entertain.* **11**(1), 5:1–5:30 (2013)
22. Hackathorn, R., Margolis, T.: Immersive analytics: building virtual data worlds for collaborative decision support. In: *Proceedings of the 2016 Workshop on Immersive Analytics (IA)*, pp. 44–47, March 2016
23. Jacoby, D., Coady, Y.: Perspective shifts in mixed reality: persuasion through collaborative gaming. In: *Proceedings of the Personalization in Persuasive Technology Workshop, Persuasive Technology 2017 (PPT 2017)*, pp. 84–90 (2017)
24. Kirsh, D.: Embodied cognition and the magical future of interaction design. *ACM Trans. Comput. Hum. Interact.* **20**(1), 3:1–3:30 (2013)
25. Klemmer, S.R., Hartmann, B., Takayama, L.: How bodies matter: five themes for interaction design. In: *Proceedings of the 6th ACM Conference on Designing Interactive systems (DIS 2006)*, pp. 140–149. ACM Press, New York (2006)
26. Lazem, S., Gračanin, D.: Social traps in Second Life. In: Debattista, K., Dickey, M., Proença, A., Santos, L.P. (eds.) *Proceedings of the 2nd International Conference on Games and Virtual Worlds for Serious Applications (VS GAMES 2010)*, pp. 133–140, 25–26 March 2010
27. Lea, R., Blackstock, M.: Smart cities: an IoT-centric approach. In: *Proceedings of the 2014 International Workshop on Web Intelligence and Smart Sensing*, pp. 12:1–12:2. ACM, New York (2014)
28. Luboschik, M., Berger, P., Staadt, O.: On spatial perception issues in augmented reality based immersive analytics. In: *Proceedings of the 2016 ACM Companion on Interactive Surfaces and Spaces*, pp. 47–53. ACM, New York (2016)
29. Matković, K., Freiler, W., Gračanin, D., Hauser, H.: ComVis: a coordinated multiple views system for prototyping new visualization technology. In: *Proceedings of the 12th International Conference on Information Visualisation (IV 2008)*, pp. 215–220, 9–11 July 2008
30. McGlenn, K., Hederman, L., Lewis, D.: SimCon: a context simulator for supporting evaluation of smart building applications when faced with uncertainty. *Pervasive Mob. Comput.* **12**, 139–159 (2014)
31. Nguyen, H., Marendy, P., Engelke, U.: Collaborative framework design for immersive analytics. In: *Proceedings of the 2016 Big Data Visual Analytics (BDVA)*, pp. 1–8, November 2016
32. Norman, D.A.: *The Design of Everyday Things*. Currency/Doubleday, New York (1990)
33. Prendinger, H., Brandherm, B., Ullrich, S.: A simulation framework for sensor-based systems in Second Life. *Presence Teleop. Virtual Environ.* **18**(6), 468–477 (2009)
34. Renambot, L., Bal, H.E., German, D., Spoelder, H.J.W.: CAVestudy: an infrastructure for computational steering in virtual reality environments. In: *Proceedings of the Ninth International Symposium on High-Performance Distributed Computing*, pp. 239–246, 1–4 August 2000
35. Roberts, J.C.: State of the art: coordinated multiple views in exploratory visualization. In: *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, pp. 61–71. IEEE, 2 July 2007

36. Sinclair, B.: *IoT Inc.: How Your Company Can Use the Internet of Things to Win in the Outcome Economy*. McGraw-Hill Education, New York (2017)
37. Thomas, J.J., Cook, K.A. (eds.): *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society, Los Alamitos (2005)
38. Williams, A., Kabisch, E., Dourish, P.: From interaction to participation: configuring space through embodied interaction. In: Beigl, M., Intille, S., Rekimoto, J., Tokuda, H. (eds.) *UbiComp 2005*. LNCS, vol. 3660, pp. 287–304. Springer, Heidelberg (2005). https://doi.org/10.1007/11551201_17
39. Wilson, M.: Six views of embodied cognition. *Psychon. Bull. Rev.* **9**(4), 625–636 (2002)
40. You, S., Thompson, C.K.: Mobile collaborative mixed reality for supporting scientific inquiry and visualization of earth science data. In: *Proceedings of the 2017 IEEE Virtual Reality Conference*, pp. 241–242, March 2017



Development and Application of the Hybrid Space App for Measuring Cognitive Focus in Hybrid Contexts

Øyvind Jøsok^{1,2(✉)}, Mathias Hedberg¹, Benjamin J. Knox^{1,5},
Kirsi Helkala¹, Stefan Sütterlin^{3,4}, and Ricardo G. Lugo²

¹ Norwegian Defence Cyber Academy, Norwegian Defence University College,
Lillehammer, Norway

{ojosok, bknox, khelkala}@fhs.mil.no,
hedberg.mathias@gmail.com

² Faculty of Social and Health Sciences, Inland Norway University
of Applied Sciences, Lillehammer, Norway

Ricardo.Lugo@inn.no

³ Faculty for Health and Welfare Sciences, Østfold University College,
Fredrikstad, Norway

Stefan.Sutterlin@hiof.no

⁴ CHTD Research Group, Division of Clinical Neuroscience,
Oslo University Hospital, Oslo, Norway

⁵ Department of Information Security and Communication Technology,
Norwegian University of Science and Technology, Gjøvik, Norway

Abstract. Insight into skills that can support agile cognitive maneuver in complex digitized contexts is necessary for improved understanding of human behaviour in contemporary society. A more digitally enabled professional life, encompassing new tools to augment reality, requires we seek further knowledge concerning which competencies make humans operate more efficiently. The Hybrid Space app was developed for collecting and analysing individual cognitive focus when engaging in hybrid contexts. This paper includes an introduction to how cognitive focus can be operationalized in The Hybrid Space conceptual framework for research purposes. It explains the development of the data collection software, The Hybrid Space app, and presents examples of data collected during a four-day cyber defence exercise at the Norwegian Cyber Defence Academy. The Hybrid Space app demonstrated ease of use for real-time analysis opportunities, as well as a reliable data collection, computation and visualization tool.

Keywords: Cyber · Software · Human computer interaction · Cognitive agility
Socio-technical system · Hybrid space · Human factors · Augmented cognition

1 Introduction

In recent years, the human factors of cyber operations and cyber security has gained increased attention [1–5]. The acknowledgment of the human as ‘the strongest link’ is more common [4, 6], due to the notion that humans remain superior to technology

when it comes to engaging in macrocognitive work (e.g. adapting to complexity through problem identification and sense making in ambiguous, shifting conditions) [7]. Humans now operate extensively in hybrid contexts, characterised by cyber and physical reciprocal determinants [8], merged with tactical and strategic level interaction [9]. More human computer interaction is placing higher demands on humans to establish cross-domain situational awareness [10] and to govern technology, whilst simultaneously complying with physical environment demands. Task demands in hybrid contexts exceed what we used to consider ‘enough’ to cope in a digitized context, and performance measures need to include soft skill proficiencies that can traverse digital and physical domains [11]. Research and understanding of the cognitive processes that support mastery of hybrid contexts are still scarce [5].

To learn more about individual cognitive manoeuvring requirements, the researchers monitored cognitive dynamics of cyber cadets in hybrid context. This paper first explains how we operationalized cognitive focus (i.e. cognitive location) by utilizing The Hybrid Space conceptual framework [9]. Then development of a self-report software, The Hybrid Space app, to help capture, visualize and analyse the cognitive focus of individuals and teams operating in hybrid contexts is presented. Further, an example describes the context in which the software was applied to capture cognitive focus of a cohort of cyber cadets at the Norwegian Defence Cyber Academy (NDCA) participating in a four-day Cyber Defence Exercise (CDX). Examples of collected data are presented and the applicability of the software is discussed.

2 Capturing Cognitive Focus

Cognitive focus can be understood as an aspect of attention that involves bringing selected information into conscious awareness [12]. Cognitive agility can be understood as the ability to be attentionally flexible, where flexible expansion and contraction of cognitive focus allows for both panoramic and selected attention in The Hybrid Space [13]. To be able to scientifically address the cognitive agility levels of personnel operating in hybrid contexts, cognitive focus has to be measured first [14]. Capturing the cognitive focus and concurrent thinking processes of individual cyber operators is a challenging task, presenting a variety of factors that can distort accuracy and validity of data. A number of methods designed to collect such data, aiming to get insight into thinking processes exists.

Behavioural task analysis is not applicable to capture cognitive focus, as this kind of information is not directly observable [15]. As a consequence, indirect techniques like visual search pattern have been utilized for tapping into the nature of expert cognitive processing in physical environments [16]. In complex domains though, experts also rely on sense making processes that couple chunks of information emerging from the cognitive domain [17]. This lead to Cognitive Task Analysis (CTA) and Cognitive Work Analysis (CWA) being developed to access cognitive elements of experts in action [18, 19]. Both are known methods to capture the cognitive processes in cyber operators [20, 21], but involve either in the moment eliciting techniques (e.g. speaking out loud while performing tasks), or retrospective methods of data collection like ‘the knowledge audit’ or ‘the simulation interview’, aimed at

uncovering cognitive demands or skills required for expert task performance [22]. While in the moment eliciting methods might be intrusive and reduce operator performance, retrospective methods might present difficulties in recalling specific cognitive focus over an extended period of time. Despite intrusiveness, CTA are proven methods for eliciting the cognitive task requirements and capturing the covert cognitive processes experts use to perform complex skills [18].

Non-intrusive techniques like functional magnetic resonance imaging (fMRI), electroencephalography (EEG) and eye-tracking are also used to assess cyber operator performance [23]. While these methods are less intrusive and can give valuable data of cognitive load and aspects of cognitive processing [24], neither give access to the kinds of cognitive focus and specific sense making processes the operator needs to accomplish in hybrid contexts.

These examples of performance data collection methods give access to important task-knowledge, cognitive load, neurological processing and what information cyber operators seek. But they do not elicit knowledge concerning how cyber operators cognitively focus and cognitively manoeuvre over time in order to make sense of the continually evolving hybrid context. Neither does it give knowledge of what skills are used to regulate cognition, nor what cognitive dynamics are beneficial to support performance in hybrid contexts. Further, in complex domains the construct of performance itself is questioned due to the complexity of interactions [2, 14, 16, 25, 26]. So far, measuring performance in cyber exercises has relied on performance measures like ‘capture the flag’ or other types of hits, errors, accuracy and time (HEAT) measurements or subject matter expert evaluation [27–29].

Embracing the full complexity of human computer interaction in hybrid contexts reveals the need to explore and measure emergent properties [30]. We argue that performance in hybrid contexts may be dependent on indicators and predictors like cognitive agility levels [14]. For now, capturing the immediate cognitive focus of personnel engaged in macrocognitive work in hybrid contexts requires self-report. In this paper, we introduce a software developed to measure cognitive focus and analyse cognitive agility.

3 The Hybrid Space Conceptual Framework as a Tool for Measuring Cognitive Focus

The Hybrid Space conceptual framework (see Fig. 1) describes the influencing factors on individual psychological conditions in hybrid contexts [9]. Hybrid refers to the complexity of interactions between agents in the cyber domain and the physical domain in this space, at all levels of hierarchy. The Hybrid Space draws attention to the human as the converging point of sense making, acknowledging the reciprocal deterministic relationship between environment, cognition and behavior, whilst acknowledging individual agency to self-influence cognitive activity [31].

The x-axis visualizes the reciprocal relationship between the physical domain on the right-hand side, and the cyber domain on the left-hand side. This subjective assessment can be seen as the cyber operator’s current cognitive focus in relation to the cyber and physical domains, and can be measured by self-reporting the location of the

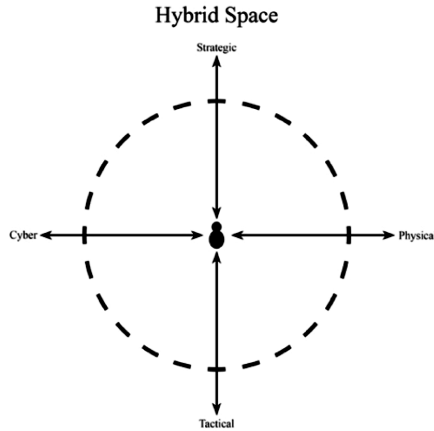


Fig. 1. The hybrid space conceptual framework [9]

momentary cognitive focus. Simultaneously sense making requirements between tactical and strategic considerations the can be marked with y-axis position. For example, when analysing malware, tactical and local considerations might be applied to sense making processes, while considering attribution might require a strategic (e.g. national or geopolitical) cognitive focus [32].

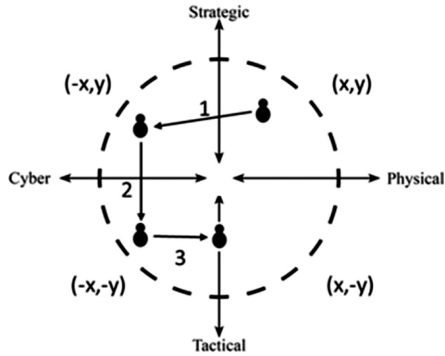
To allow for measurement of cognitive focus using The Hybrid Space app, the conceptual framework is mapped in a Cartesian plane. The x and y axis movements are limited to -100 to $+100$. Based on the cognitive focus of each movement, The Hybrid Space is operationalized through four indices (See Table 1, and Fig. 2). These indices are quantitative proxies for the construct of cognitive agility, and can be utilized as dependent variables in the case of further statistical analysis [14]. When developing The Hybrid Space app, two sliders and a text window were added to allow for collection of additional data of interest. The sliders are comprised of a 9-point Likert scale (See. Fig. 3), and the slider text can easily be changed at the researchers’ prerogative.

Table 1. Explanation of indices

HSDT	Distance traveled in the Cartesian Plane measured by Euclidean distance
HSQC	Number of quadrant changes
HSxM	Movement along the cyber-physical domain (x-axis)
HSyM	Movement along the strategic-tactical domain (y-axis)

4 Software Description – The Hybrid Space App

To develop the software, a spiral lifecycle methodology was used. This allows for continuous feedback on the development process. This involved creating an initial stripped-down prototype of the software based on the research requirements. The initial prototype was then tested to identifying strengths, weaknesses and risks with the



Movement in the Hybrid Space: (1) operator reporting quadrant change (x,y) to $(-x,y)$; (2) operator moving along the y-axis; (3) an operator reporting movement to an axis but not crossing to other quadrant

Fig. 2. Operationalization of the hybrid space movements [14]

The Hybrid Space

Control: 5

Effort: 5

Comment

Submit

Fig. 3. The hybrid space app user interface

current prototype. Changes were then proposed for the next revision of the prototype, and a new revision of the prototype created. This process ensured that a final product could be created in collaboration between researchers and programmer.

The software is based on three main components; a backend database component, a database interaction component and a front-end component. The backend database is where all the collected data is stored, in addition to also being where all the business logic of the system is processed. This component is based on Postgres (PostgreSQL), an open source object-relational database management system. Postgres handles all equations used and calculations done by The Hybrid Space app through a set of triggers that are activated when new data is added to the database. The database structure and triggers are defined in a simple database initialization script that Postgres processes upon initialization of the database. This file can be located at `/Docker/hybridspace_db/initdb.sql` in the source code.

The backend interaction component allows for user interaction with the database. This component is built using a web application framework known as ExpressJS. The component handles all queries to the database and is a RESTful service, listening for requests over HTTP and replying with JSON data. Creating a RESTful service standardizes the way users and other applications interact with the system. Some examples are shown in Table 2.

Table 2. Hybrid Space REST API examples:

<p>Get evaluations for person 1:</p> <pre>\$ curl http://127.0.0.1:3000/api/search?person=1 {"status":"success","data":[{"evalid":1,"person":1,"team":"Lag 1", ...}</pre>
<p>Add evaluation for person 1:</p> <pre>\$ curl --data "person=1&x=22&y=3&slider1=2&slider2=3&comment=test2&password=password1" http://127.0.0.1:3000/api/evaluation</pre>

This is currently the only way to add, modify or delete data in the database. The query possibilities can be found in the api routes definition file found in `/routes/api.js`. The functions and their associated database queries can be found in the file `queries.js`, located in the project root directory.

The last component of the system is the front end. This component gives the user an interface to make REST queries to the server, and parse the JSON data the server replies with. The current front end is web based, using JavaScript to send and parse data. In the future, this component could be based on a native Android or iOS application due to the flexibility associated with using a REST API for server communication.

When new data is added to the system, a set of calculations are done. The distance traveled (HSDT) is calculated using the movement along the x -axis (HSxM) and y -axis (HSyM). HSxM and HSyM are calculated by finding the absolute value between the new plot x/y value and the existing x/y plot found in the database. The HSDT value is then calculated using basic trigonometrics:

$$\text{HSDT} = \sum_{i=1}^n \sqrt{(\text{HSxMi})^2 + (\text{HSyMi})^2}$$

This calculation can be found in the function 'trg_travel' function located in the `initdb.sql` file. A snippet of the code performing the logic can be found in Table 3.

Table 3. Calculation of distance

```
xtrav := (@((SELECT x FROM evaluations WHERE pid =
NEW.pid ORDER BY evalid DESC LIMIT 1 ) - NEW.x));
ytrav := (@((SELECT y FROM evaluations WHERE pid =
NEW.pid ORDER BY evalid DESC LIMIT 1 ) - NEW.y));
NEW.xtravel := xtrav;
NEW.ytravel := ytrav;
NEW.travel := round(|/ ((xtrav^2) + (ytrav^2)));
```

The system also keeps track of what quadrant the user resides in so that any quadrant changes can be registered. This calculation is done using the function 'trg_quadchange' located in the `initdb.sql` file. A snippet of the code performing the logic can be seen in Table 4.

Table 4. Calculation of quadrant change

```
oldquad := (SELECT quad FROM evaluations WHERE pid =
NEW.pid ORDER BY evalid DESC LIMIT 1 );
IF NEW.quad = oldquad THEN
    NEW.quadchange := FALSE;
ELSE
    NEW.quadchange := TRUE;
END IF;
```

This data allows for more advanced queries to the system, such as how many quadrant changes a team has, or how much travel a specific person has.

4.1 Installation

The software was developed to run in a Docker container, however it is also possible to install the software directly on a Linux host without any containerization. Currently, the software has only been tested on Fedora 24/25/26, however other operating systems that support Docker should function just fine. Docker is supported on most Linux distributions, with limited support for Windows hosts. It is therefore recommended to run the software on a Linux distribution such as Fedora, CentOS or Ubuntu.

As mentioned earlier, it is possible to run the system directly on a Linux host without any containerization, however the authors recommend using Docker as it simplifies the installation process significantly. The Hybrid Space wiki documentation included with the source code details this process. In short, this process involves installing `nodejs` and `postgresql` on a server, setting up the database access and running the application.

The Hybrid Space app is freely available for download at <http://github.com/metrafonic/TheHybridSpace>. Guides to installing and using can be found in the connected Wiki.

5 Collection of Cognitive Focus During a Cyber Defence Exercise

Collection of cognitive focus was performed during the annual CDX at the NDCA, November 2017. A complete cohort undergoing a cyber engineer education totalling 38 cyber cadets participated in the CDX. They worked in four independent student teams (one team consisted of 9 or 10 members) in four separate rooms. 23 cadets participated in this research using The Hybrid Space app.

All of the participants were subject to the same external cyber and physical activity, framed in the same scenario. Ongoing momentary assessment of perceived cognitive focus was conducted as cadets were instructed to mark their cognitive location in The Hybrid Space around every full hour (0800–1800). Simultaneously the perceived level of control and cognitive effort were assessed by adjusting the sliders (see Fig. 3). Comments were made voluntarily in order to minimize intervention time. The assessments were repeated for four consecutive days throughout the course of the CDX, giving a total of 854 data entries in The Hybrid Space app.

The CDX is constructed to challenge participants to apply cognitive agility to make sense of the multi-domain reciprocal dependencies in The Hybrid Space. This is achieved by using a scenario where both strategic situation information and tactical information is injected as part of a holistic scenario driven events matrix. In addition, cadets are challenged to communicate their current multi-domain situational awareness in relation to the overall evolving scenario.

6 Examples of Research Results with the Hybrid Space App

During the course of the four day CDX researchers were able to monitor the entry of data to ensure adherence to instructions and reduce the possibility of missing values. The Hybrid Space app graphical output is shown in Fig. 4 with the complete movement of one person during the four days, as well as variations in control and effort easily accessible and visualized. In a real-time analysis, this gives access to individual cognitive focus, where participants struggle, have control or where they put in effort. Compared with the Exercise Control (EXCON) knowledge of scenario developments and task requirements this data can shed light on how participants manoeuvre and focus to make sense of information emerging from cyber and physical domains.

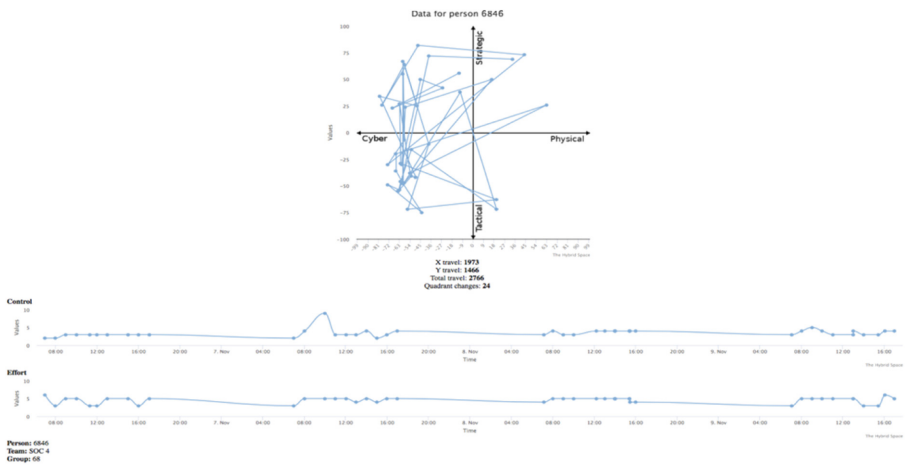


Fig. 4. Example of data collected with The Hybrid Space app

Each data entry comprises of an automatically generated ID number, username, team name and time stamp. User entries comprised of x location, y location, control, effort and voluntary comment. Raw data backend is shown in Fig. 5.

An additional view, Fig. 6, showing the aggregated number of data entries of each person and each team was devised to give an easier overview of the data collection process, as well as the possibility to export data on individuals and/or team. This reduces data handling from raw data as well as provides an overview that can be useful when comparing the data on team or group level.

In contrast to paper-pencil solutions that we have used in early stages of cognitive agility research, The Hybrid Space app proved to be less invasive and removed manual data transfer errors. During the period of data collection, preliminary data analysis can be conducted in real-time, and the cognitive dynamics can be visualized instantly.

Evaluations Live •

Status: OK

827
 Evaluations

Add evaluation

Applied Filters:

Add Filters:

Apply filter to the view:
 =

evalid	person	team	collection	x	y	travel	slider1	slider1text	slider2	slider2text	comment	time	
64	6845	SOC 4	68	34	34		7	Control	6	Effort		2017-11-06T06:55:28.042Z	Delete
65	6822	SOC 2	68	95	1		9	Control	1	Effort		2017-11-06T06:57:45.944Z	Delete
66	6824	SOC 2	68	62	33		5	Control	5	Effort		2017-11-06T06:58:22.588Z	Delete
67	6827	SOC 2	68	91	17		5	Control	5	Effort		2017-11-06T06:58:42.995Z	Delete
68	6816	SOC 1	68	24	71		2	Control	5	Effort		2017-11-06T06:59:08.959Z	Delete
69	6846	SOC 4	68	34	69		2	Control	6	Effort		2017-11-06T06:59:10.047Z	Delete
70	6842	SOC 4	68	-35	-43		6	Control	5	Effort		2017-11-06T06:59:11.881Z	Delete
71	6829	SOC 2	68	-21	-23		5	Control	5	Effort		2017-11-06T06:59:44.627Z	Delete
72	6840	SOC 4	68	0	-45		2	Control	3	Effort		2017-11-06T06:59:56.634Z	Delete
73	6818	SOC 1	68	-1	43		3	Control	3	Effort	Morgenbrief	2017-11-06T07:00:22.597Z	Delete
74	6834	SOC 3	68	-1	0		1	Control	1	Effort		2017-11-06T07:02:01.106Z	Delete
75	6844	SOC 4	68	4	0		2	Control	1	Effort		2017-11-06T07:09:24.441Z	Delete
76	6848	SOC 4	68	33	0		7	Control	3	Effort		2017-11-06T07:11:18.326Z	Delete
77	6843	SOC 4	68	46	0		9	Control	3	Effort		2017-11-06T07:12:07.969Z	Delete
78	6849	SOC 4	68	-35	58		4	Control	5	Effort		2017-11-06T07:14:27.033Z	Delete

Next
[Download CSV data](#) | [View JSON data](#)

Fig. 5. Data entries administrator view

Persons Live •

Status: OK

38
 Registered persons

All Persons

The PID is a unique identifier for each person amongst all the datasets. This number is only to be used when deleting persons.
 The Person number is what the person should be referred to. When they log in, they must use this number

pid	person	team	collection	password	evaluations	quadschanges	xtravel	ytravel	travel	
10	6811	SOC 1	68	CDX6811	38	22	1723	1853	2854	View
11	6812	SOC 1	68	CDX6812	6	6	541	477	724	View
12	6813	SOC 1	68	CDX6813	6	5	210	429	535	View
13	6814	SOC 1	68	CDX6814	18	6	456	446	723	View
14	6815	SOC 68	68	CDX6815	0	0				View
15	6816	SOC 1	68	CDX6816	38	14	642	866	1180	View
16	6817	SOC 1	68	CDX6817	36	12	791	763	1232	View
17	6818	SOC 1	68	CDX6818	26	18	1869	1369	2602	View
18	6819	SOC 1	68	CDX6819	41	13	1401	1785	2516	View
19	6821	SOC 2	68	CDX6821	0	0				View

All Teams

team	collection	members	evaluations	quadschanges	xtravel	ytravel	travel
SOC 1	68	9	209	96	7633	7988	12366
SOC 2	68	9	158	72	8152	4899	10305
SOC 3	68	10	87	45	3313	3588	5504
SOC 4	68	10	373	209	17639	14269	25195

[Download CSV data](#)

Add person

When adding users, make sure that every user has a unique person number. This will be used to login.

Person number	Team	Group	Password
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Delete person

WARNING:
 Make sure to use the PID, not the person number. The PID can be found in the user table.

PID

Troubleshooting:
 We are only able to delete users that have no data associated with them. If there is no response, go to the evaluations

Fig. 6. Data entries team and group administrator view

7 Applicable Contexts and Further Development

The application of The Hybrid Space app in combination with The Hybrid Space conceptual framework evaluations in the context of education and training of cyber cadets can give insight into so far unexplored cognitive dynamics on individual and group level. From an EXCON perspective, the data might be useful for observing cognitive focus during the course of a CDX. The software can also be used for debriefing and/or as a

complimentary tool for conducting CTA after a training session is completed. Further research can be conducted applying statistical analysis of The Hybrid Space movements with data from other inventories measuring i.e. self-regulation, metacognitive awareness or other constructs known to support cross domain performance [13]. Such combination of data can shed light on beneficial cognitive traits and competencies supporting agile manoeuvre in The Hybrid Space [13, 14]. Over time, and in combination with valid performance measures, further research can produce knowledge about cognitive skills that are beneficial for operating in hybrid contexts.

The visual representation given in the administrator view (Fig. 4) gives an overview of the predominant cognitive focus at an individual level (i.e. direct access to a part of the cognitive dynamics of each participant). With further development, this could also be expanded to visualize predominant cognitive focus at various group levels (e.g. team and unit). As cyber operator performance is dependent on both individual (metacognitive) and team (macrocognitive) work [9, 11, 19], collection of such data could contribute to develop knowledge and understanding of beneficial cognitive focus and dynamics displayed by cyber operators in hybrid contexts. However, using The Hybrid Space for research purposes demands a level of prior understanding among researcher and research objects. Pathways for improved understanding of hybrid contexts among cyber cadets is a developing area of research [11, 13].

While The Hybrid Space app is developed utilizing the military cyber operator practice, there are opportunities to adopt the framework, and the software, to research other digitally mediated cyber-physical contexts where humans operate in and through digital technology. An example could be to research how digitally mediated learning platforms connected to cyberspace influence the classroom context and impacts teacher-student power gradient and dynamics. Or to research the cognitive dynamics displayed by parents and children in the family practice as they are influenced by more time spent in the world of cyber.

8 Conclusion

The Hybrid Space app is a software tool providing the researcher with a developed software and method of capturing and visualizing momentary cognitive focus and the dynamics of individuals in hybrid contexts. Compared to other methods of cognitive data collection like CTA, fMRI or EEG, using The Hybrid Space app gives access to new and qualitatively different data on individual cognitive dynamics with a minimum of intrusion. The software further provides the opportunity to visualize cognitive dynamics for research, teaching and presentation analysis. The Hybrid Space app provides necessary computation options of variables and displays various measures of movement in The Hybrid Space, on individual and group level. It also provides access to easy research setup and use, due to its graphical interface and versatile platform possibilities. Finally, researchers are provided with access to export all collected data with timestamps in comma-separated values (CSV) format. The Hybrid Space software provides a graphical user interface that makes it applicable for both research and

teaching purposes. An open access approach and compatibility allows further development and/or integration with other software tools maximizing benefits by having possibility of combining various software solutions.

References

1. Gutzwiller, R.S., Fugate, S., Sawyer, B.D., Hancock, P.: The human factors of cyber network defense. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 322–326. SAGE Publications (2015)
2. Mancuso, V.F., Christensen, J.C., Cowley, J., Finomore, V., Gonzalez, C., Knott, B.: Human factors in cyber warfare II emerging perspectives. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 415–418. SAGE Publications, (2014)
3. Knott, B.A., Mancuso, V.F., Bennett, K., Finomore, V., McNeese, M., McKneely, J.A., Beecher, M.: Human factors in cyber warfare: alternative perspectives. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 399–403. SAGE Publications, Los Angeles (2013)
4. Helkala, K., Knox, B.J., Jøsok, Ø., Lugo, R.G., Sütterlin, S., Dyrkolbotn, G.O., Svendsen, N.K.: Supporting the human in cyber defence. In: Katsikas, S.K., Cuppens, F., Cuppens, N., Lambrinouidakis, C., Kalloniatis, C., Mylopoulos, J., Antón, A., Gritzalis, S. (eds.) CyberICPS/SECPRE -2017. LNCS, vol. 10683, pp. 147–162. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-72817-9_10
5. Ben-Asher, N., Gonzalez, C.: Effects of cyber security knowledge on attack detection. *Comput. Hum. Behav.* **48**, 51–61 (2015)
6. D’Amico, A., Buchanan, L., Kirkpatrick, D., Walczak, P.: Cyber operator perspectives on security visualization. In: Nicholson, D. (ed.) *Advances in Human Factors in Cybersecurity. Advances in Intelligent Systems and Computing*, vol. 501, pp. 69–81. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41932-9_7
7. Klein, G., Ross, K.G., Moon, B.M., Klein, D.E., Hoffman, R.R., Hollnagel, E.: *Macro-cognition*. *IEEE Intell. Syst.* **18**, 81–85 (2003)
8. Pfleeger, S.L., Caputo, D.D.: Leveraging behavioral science to mitigate cyber security risk. *Comput. Secur.* **31**, 597–611 (2012)
9. Jøsok, Ø., Knox, Benjamin J., Helkala, K., Lugo, Ricardo G., Sütterlin, S., Ward, P.: Exploring the hybrid space. In: Schmorow, D., Fidopiastis, C. (eds.) *AC 2016. LNCS (LNAI)*, vol. 9744, pp. 178–188. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39952-2_18
10. Franke, U., Brynielsson, J.: Cyber situational awareness—a systematic review of the literature. *Comput. Secur.* **46**, 18–31 (2014)
11. Knox, B.J., Jøsok, Ø., Helkala, K., Khooshabeh, P., Ødegaard, T., Kwei-Narh, P., Lugo, R. G., Sütterlin, S.: Socio-technical communication: The Hybrid Space and the OLB-Model for science-based cyber education. *J. Mil. Psychol.* (Submitted)
12. MacKay-Brandt, A.: Focused Attention. In: Kreutzer, J.S., DeLuca, J., Caplan, B. (eds.) *Encyclopedia of Clinical Neuropsychology*, pp. 1066–1067. Springer, New York (2011)
13. Knox, B.J., Lugo, R.G., Helkala, K., Sütterlin, S., Jøsok, Ø.: Education for cognitive agility: improved understanding and governance of cyberpower. In: *17th European Conference on Cyber Warfare and Security* (2018)

14. Knox, B.J., Lugo, R.G., Jøsok, Ø., Helkala, K., Sütterlin, S.: Towards a cognitive agility index: the role of metacognition in human computer interaction. In: Stephanidis, C. (ed.) HCI 2017. CCIS, vol. 713, pp. 330–338. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58750-9_46
15. Clark, R.: Cognitive task analysis for expert-based instruction in healthcare. In: Spector, J. M., Merrill, M.D., Elen, J., Bishop, M.J. (eds.) Handbook of Research on Educational Communications and Technology, pp. 541–551. Springer, New York, New York, NY (2014). https://doi.org/10.1007/978-1-4614-3185-5_42
16. Ward, P., Suss, J., Basevitch, I.: Expertise and expert performance-based training (ExPerT) in complex domains. In: Technology, Instruction, Cognition and Learning, vol. 7, pp. 121–145 (2009)
17. Ward, P., Ericsson, K.A., Williams, A.M.: Complex perceptual-cognitive expertise in a simulated task environment. *J. Cogn. Eng. Decis. Making* **7**, 231–254 (2013)
18. Clark, R.E., Feldon, D., vanMerriënboer, J., Yates, K., Early, S.: Cognitive task analysis. In: Spector, J.M., Merrill, M.D., van Merriënboer, J.J.G., Driscoll, M.P. (eds.) Handbook of research on educational communications and technology, vol. 3. Lawrence Erlbaum Associates, Mahwah (2008)
19. Lathrop, S.D., Trent, S., Hoffman, R.: Applying human factors research towards cyberspace operations: a practitioner’s perspective. In: Nicholson, D. (ed.) Advances in Human Factors in Cybersecurity: Proceedings of the AHFE 2016 International Conference on Human Factors in Cybersecurity, 27–31 July 2016, Walt Disney World®, Florida, USA, pp. 281–293. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41932-9_23
20. Champion, M., Jariwala, S., Ward, P., Cooke, N.J.: Using cognitive task analysis to investigate the contribution of informal education to developing cyber security expertise. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 58, pp. 310–314 (2014)
21. Champion, M.A., Rajivan, P., Cooke, N.J., Jariwala, S.: Team-based cyber defense analysis. In: 2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, pp. 218–221. IEEE (2012)
22. Militello, L.G., Hutton, R.J.B.: Applied cognitive task analysis (ACTA): a practitioner’s toolkit for understanding cognitive task demands. *Ergonomics* **41**, 1618–1641 (1998)
23. Edgar, T.W., Manz, D.O.: Research Methods in Cyber Security (2017)
24. Eckstein, M.K., Guerra-Carrillo, B., Miller Singley, A.T., Bunge, S.A.: Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Dev. Cogn. Neurosci.* **25**, 69–91 (2017)
25. Jøsok, Ø., Knox, Benjamin J., Helkala, K., Wilson, K., Sütterlin, S., Lugo, Ricardo G., Ødegaard, T.: Macrocognition applied to the hybrid space: team environment, functions and processes in cyber operations. In: Schmorrow, Dylan D., Fidopiastis, Cali M. (eds.) AC 2017. LNCS (LNAI), vol. 10285, pp. 486–500. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58625-0_35
26. McNeese, M., Cooke, N.J., D’Amico, A., Endsley, M.R., Gonzalez, C., Roth, E., Salas, E.: Perspectives on the role of cognition in cyber security. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 268–271. SAGE Publications, Los Angeles (2012)
27. Hoffman, R.R., Patterson, E., Miller, J.: Some challenges for macrocognitive measurement. In: Macrocognition Metrics and Scenarios: Design and Evaluation for Real-World Teams, pp. 11–28 (2009)
28. Buchler, N., Fitzhugh, S.M., Marusich, L.R., Ungvarsky, D.M., Lebiere, C., Gonzalez, C.: Mission command in the age of network-enabled operations: social network analysis of information sharing and situation awareness. *Front. Psychol.* **7**, 937 (2016)

29. Forsythe, C., Silva, A., Stevens-Adams, S., Bradshaw, J.: Human dimension in cyber operations research and development priorities. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2013. LNCS (LNAI), vol. 8027, pp. 418–422. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39454-6_44
30. Hoffman, R.R., Hancock, P.A.: Measuring resilience. *Hum. Factors* **59**, 564–581 (2017)
31. Bandura, A.: *Social Foundations of Thought and Action - A Social Cognitive Theory*. Prentice Hall, New Jersey (1986)
32. Gibney, A.: *Zero Days. World War 3.0*. Magnolia Pictures, United States (2016)



Identifying Affordance Features in Virtual Reality: How Do Virtual Reality Games Reinforce User Experience?

Jumin Lee¹, Jounghae Bang^{2(✉)}, and Hyunju Suh³

¹ Kyung Hee Cyber University, #26 Kyungheedaero, Dongdaemun-gu, Seoul, Korea

² Kookmin University, 77 Jeongneung-ro, Seoul, Korea
bangjh@kookmin.ac.kr

³ Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul, Korea

Abstract. In order to clarify the concepts of affordance, we reviewed the previous studies of presence, interactivity, and affordance. Then, we presented the three-step identifying affordance features (IAF) model, which were (1) to develop a stage-of-the-action model for VR game environment that defined behaviors in the VR space from the user's point of view, (2) to identify affordance features based on the stage-of-action model, and (3) to identify dimensions and to ensure the affordance features with the two-dimension framework for VR games. The results of this study will provide a consistent and comprehensive academic approach to the role of affordance to enhance user experience in VR games. Also, it is expected that VR devices and applications will become a basis for understanding user-oriented design and service factors in the process of developing and distributing VR devices and applications.

Keywords: Virtual reality · Virtual reality games · Affordance
Presence · Interactivity

1 Introduction

The VR market was expected to grow at an annual average rate of 80% or more, reaching \$ 40 billion by 2020 [39]. In particular, after 2018, VR applications were expected to exceed the hardware market size, and game content was expected to account for the largest portion at the beginning. As the prices of equipment dropped and the number of VR experience centers increased, the market seemed to expand rapidly and the more investment has been made. However, the market did not expand as expected. Market researcher Digi-Capital [8] predicts that the VR market will be worth \$ 25 billion by 2021 from Augmented/Virtual Reality Report 2017. This is a significant downward revision from the April 2015 report. The company had expected the VR market to reach \$ 30 billion by 2020.

VR applications provide users with the opportunity to experience a virtual presence through 3D computer graphics that react to user movement [36]. At the present time, there are insufficient virtual reality applications other than visual elements [27, 36], and unidirectional delivery of defined content rather than interaction. Due to these

limitations, users do not experience sufficient affordance, and such restriction makes obstacles against users' virtual experience. In the end, it becomes a stumbling block to reuse VR application and to grow the VR industry.

In order to expand the market, we need more realistic and easy-to-use content that induce users who is willing to pay VR content in the early stage of the industry. Although VR market is growing, research on VR application still remained in conceptual stage base on reference reviews, experiments, and focus-group interviews without integrated framework. To overcome the limitation of previous research, this study proposed the three-step conceptual framework to identify right affordance features in the integrated point of view, which are key factors to enhance presence in VR and therefore strengthen users' virtual experience. We have focused on head mounted display (HMD)-based VR game environment which is state-of-the-art at present.

In order to clarify the concepts of affordance, we first reviewed previous studies of presence, interactivity, and affordance. Then, we presented the three-step identifying affordance features (IAF) model, which were (1) to develop a stage-of-the-action model for VR game environment that defined behaviors in the VR space from the user's point of view, (2) to identify affordance features based on the-stage-of-action model, and (3) to identify dimensions and to ensure the affordance features with the two-dimension framework for VR games. The results of this study will provide a consistent and comprehensive academic approach to the role of affordance to enhance user experience in VR games. Also, it is expected that VR devices and applications will become a basis for understanding user-oriented design and service factors in the process of developing and distributing VR devices and applications.

2 Literature Review

2.1 Presence in VR Game

When users put on HMD, they feel presence in virtual world. Presence was simply defined as the perception that nothing is between one's self and the virtual world. In that sense, Lombard and Ditton [24] said presence as the perceptual illusion of non-mediation. Lee [21] proposed that the presence was a psychological state in which virtual objects are experience as actual objects in either sensory or non-sensory ways. Presence is frequently presented as consisting of two phenomena: spatial presence and social presence [3]. Spatial presence is defined as the sense of "being there" including automatic responses to spatial cues and the mental models of mediated spaces that create the illusion of place [3, 4]. Spatial presence is influenced by technological determinants and user-based determinants. Typical technological determinants of spatial presence are the degree of interactivity of a mediated spatial environment, the breadth of human sensory channels addressed by the environment, and the naturalness of proved spatial information across sensory channels [3, 37]. Therefore, when you play a VR game, 360-degree body movements, perspective and 3D objects will help you feel spatial presence. With the technological determinants, typical user-based determinants are a person's interest in and attention to the mediated spatial environment, user's arousal level [2] and his or her cognitive spatial ability [41]. The user-based determinants are

difficult to handle by technology due to individual differences. However VR games convey additional information in a more visually or audibly emphasized way to increase arousal level and cognitive spatial ability.

Social presence is defined as a “sense of being together with another,” including primitive responses to social cues, simulations of “other minds,” and automatically generated models of the intentionality of others (people, animals, agents, gods, and so on). Social presence is studied to explore some aspects of technology or the effects of technology. Researchers in communication and human-computer interaction area are typically interested in social presence because it may mediate the effects of other central variables such as attitudes towards the mediated others, features of the interface, persuasion, illusions of reality, learning and memory, and mental health [1, 4]. Social presence in VR game include VR interface including game navigation and interaction with other things/persons/users. This social presence helps users enjoying and immersing the VR game.

2.2 Interactivity in VR Games

The interactivity of game system is known as an essential element of video games due to its influences on the uses of the video games [40] as well as of presence experiences [18]. Interactivity was defined as the degree to which a user can actively participate in certain experience by controlling the forms and/or contents [19, 37]. As a psychological variable, the concept of interactivity was developed based on two types of efficacy, which were internal and external [32]. That is, if interactivity can be perceived, one person can send a message to a receiver, who can send the feedback to the sender. In the online environment, the internal efficacy can be seen as users’ perceived control over where they are and where they are going while external efficacy is viewed as “externally based system efficacy” which is users’ sense of how responsive a Web site is to the users’ actions [42].

As studies on interactivity in the new technology, information systems or new media, more dimensions of interactivity have been explored. Therefore this view of perceived interactivity as two dimensions has been extended to the arguments in which interactivity includes three dimensions, which are control (internal efficacy), responsiveness (external system efficacy) and communication (direction) [23, 30], or more dimensions (features) such as time sensitivity [29], speed of feedback [7], and complexity of choice available [15]. It is found that the very basic and common dimensions of interactivity are control and responsiveness.

Previous research noted that interactivity with the perception of control is one of the factors that boost the enjoyment of video games [17, 40]. Moreover, a video game gets many different people from many different places networked via online and engaged in the game. Therefore their plays are all different and unpredictable, which cause various random situations in the game and all the attentions from the players [20].

Even for the HMD-based VR games, which are up-to-dated format of video games, perceived interactivity is an important factor to feel presence. VR environment with HMD provides ‘a 100% of virtual space’ and the opportunity to experience a virtual presence, and therefore users can immerse into the space and feel presence [36].

Because of these features of VR, it is critical to provide perceived interactivity and presence, for which affordance becomes essential.

2.3 Affordance Theory

VR game should provide affordances that feel the presence of virtual world and affordances that allow you to immerse the game. Affordance describes the physical interaction between objects and users. Gibson [12], an ecological psychologist who first introduced the concept of Affordance, defined affordance as “everything that a human environment provides and stimulates.” It means that the physical relationship between the user and the object reflects what is possible in the object, such as what things look like. For example, ‘a knee-high object with a hard, horizontal, wide surface’ is an object with affordance that induces the action ‘sit’. The user perceives and behaves by observing the information of the subject without previous knowledge or instructions. That is, Gibson’s affordance is a fixed characteristic and capability of the object [12].

On the other hand, Norman [33], who approaches affordance in terms of the perception of the user rather than an independent feature of the environment, describes affordance as “the perceived characteristics of things, or the actual characteristics of things.” He distinguished *perceived affordance* with Gibson’s *real affordance*. The interpretation of the object from each user’s experience does not necessarily match the intended property or design of the object. His definition provided a foundation for attention to the affordance concept in mediated communication theories, especially human computer interaction (HCI) field, which studies the interaction between humans and computers,

In contrast to researchers who succeeded Gibson [12] and found affordance in fixed capabilities and features of the object [11, 28], Hartson [14] focused on what reaction process the user takes to relate to things in interaction design and extended and supplemented the perceived affordance concept of Norman [33]. He stated that “the concept of affordance is an instrument that focuses on the link between user, action, and design. The process of recognizing and acting from what the user feels shows how to learn and use things by each of the absences [14].” In other words, the reason for the importance of affordance is that the physical characteristics of the tool cause intuitive behavior of the user without using a high level of cognitive processing such as inference or prior knowledge when the user uses the tool.

He renamed Norman’s perceived affordance [33] as *cognitive affordance* and real affordance as *physical affordance*. Cognitive affordance, also referred as perceptual information about (real) affordance helps users to think and/or know about the object and physical affordance facilitates users to do something physically. “Clear and precise words in a button label [14]” could be a cognitive affordance feature because it supports users to understand the function of the button. Reasonable size and accessible location of the button could be a physical affordance feature because it helps users to click the button easily. He added two more affordance concepts: *sensory affordance* and *functional affordance*. Sensory affordance is a design feature about user’s sensory experience and includes some lexical and syntactic interpretation, but not about semantic interpretation. In addition, Hartson [14] included purpose in the definition of physical affordance and

suggested the concept of functional affordance. In terms of Human-computer interface (HCI), you can click anywhere on the screen but do not click just because it is possible. Users click buttons to accomplish a goal and systems will respond to the action. For example, a door is a physical affordance because it “can be grasped and turned”, and also a functional affordance because it can be “grasped and turned in order to operate the door (that allows to pass).” In addition, Hartson [14] referred to Norman’s ‘Stage-of-Action’ model [34] and explain the process which users interact with some machine.

2.4 Affordance in Application Area

In their social media affordance study to support continuous communication in a social media environment, Majchrzak et al. [25] found four affordances to engage in dialogue for knowledge sharing: *metavoicing*, *triggered attending*, *network-informed associating*, and *generative role-taking*. The research focused on engaging and associating relationships for online activities. In order to succeed the results of this study and to derive the affordance in the VR game environment, the technical characteristics of the VR different from the social media should be additionally considered. In the same vein, Lee and Shin [22] analyzed the product design and multi-media affordance cases of mobile games with high game rankings and *identified interaction*, *user experience*, *metaphor* and *simplicity* as the affordance of mobile game.

Although new concepts such as technology affordance [11], communication affordance [16], pedagogical affordance [38] and social affordance [10, 25] have emerged as a result of studying affordance in various fields, most studies that explored the interaction between users and computerized beings in augmented reality games and virtual reality games are mostly dependent on concepts Hartson [14] suggested.

The concept of affordance used in virtual reality research for a long time, but is still in the early stage of theory development based on specific products and environment [13, 31]. Each of the studies have different definition and classification of affordance concept [6, 9, 26]. A generally accepted definition of affordance and the relationship among diverse affordance concepts in VR game environment is needed to develop more sophisticated theory and research models for empirical studies.

3 Theoretical Framework for Affordance Development in VR Games

3.1 Three-Step IAF (Identifying Affordance Features) Model

To enhance users’ experience, affordance are very basic key concepts and therefore it is important to identify affordance features for specific industry (here is VR games) to increase presence. Here this study presents the three-step IAF model. The three steps are (1) to develop a stage-of-the-action model for VR game environment that defines behaviors in the VR space from the user’s point of view, (2) to identify affordance features based on the-stage-of-action model, and (3) to identify dimensions and to ensure the affordance features with the two-dimension framework for VR games.

3.2 Step 1: The Stage-of-Action Model of VR Games

Affordance is the first step in cognitively letting a user feel presence, technically being the starting point for making VR games sophisticated. In this regard, studies on the user's feeling of affordance in the virtual space under the VR environment [35] and on the significance of affordance in the degree of presence [13] have been done. The preconditions to understand the role of affordance in creating and maintaining a virtual reality is to believe that there is an alternative "place" where we can presence. When the user's viewpoint changes from observer to actor, it increases the feeling of "place". This means not only believing in the space the user can explore, but also feeling presence as if they are one of the objects in this space. Users perform actions in an environment full of tools to achieve goals. The time and space of action is limited by the environment and the affordance of the object [13].

Based on the concept of affordance and presence in the VR game environment discussed above, we have modified and improved the stage-of-action model of Norman [34] to fit the VR environment. The Norman [34] model is a chronological summary of the typical user activity that occurs when a user interacts with a machine ("some machine").

In the VR game environment, user behavior has different aspects from the case of using a general machine which Norman [34] suggested. First, an action takes place in a space separated from the real world as a way to enter 3D virtual space while a user act in the real world and the action applied into two-dimension monitor in a general machine. Second, movement in the virtual space is done not only through the manipulation of the device but also through user's movement. Third, the game is played through the player's initiative action and interaction with various objects. Especially for RPG games social factors should be considered because many people from different places log in to the game and play together. Forth, the user is continuously exposed to a plurality of various missions and feed-backs of the results to perform the next successive mission. Based on the above analysis, we derive a new state-of-action model in VR environment. Users will encounter affordance and presence on the process of experiencing VR games.

As seen in Fig. 1, gamers enter the virtual world when they wear HMD. After they put on HMD and see the virtual world, they perceive their presence in the virtual world (Perceiving Virtual World). Involuntary attention should be premised on the spatial presence. Various sensory factors and depth of presented information such as real image is critical. Second, in order to play the game, users should perceive every movement and line of sight as first-person position (Perceiving First-Person Position). When they go close to the object in virtual world, the objects should be close-up. When they turn their face and move gaze, the virtual world is changed along their eyes.

Third, the gamers recognize what they are doing based on beings and things around in virtual world (Interpreting the Perception). When they perceive virtual world and the role in the game, they decide to what to do in the game (Goal Setting). Goal setting stage is the beginning point of the real game. When they execute the actions such as moving, click, shaking, or stand up/down, cognitive, physical, and sensory affordance can support to execute in VR game. After the action, gamers see the results of the action. Physical and sensory affordance helps them to understand the feedback and

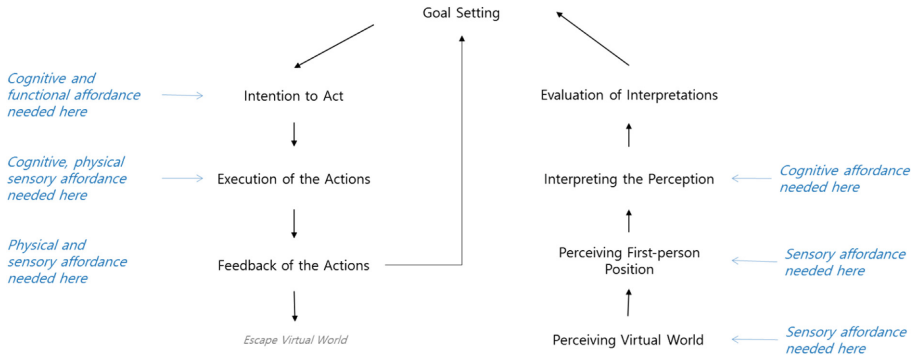


Fig. 1. Stage- of-action model for VR game environment (draft)

recognize the next step of the game. The loop from Goal Setting to Feedback of the actions continues before they stop the game and escape the virtual world.

In addition, we mapped four types of affordance, i.e. cognitive affordance, physical affordance, sensory affordance, and functional affordance from Hartson [14], to our stage-of-action model. The range of each affordance is extended in the VR environment. In cognitive affordance perspective, users are exposed to unfamiliar and diverse stimuli and must perform missions within them. Therefore, the cognitive affordance feature should be enhanced to guide the journey to perform the goal in the VR environment. In terms of physical affordance, the physical features of action itself and the results of actions should be proper and reasonable. The responses of VR world, background 3D image as well as mission-related images, should interact as expected to strengthen the reality of virtual environment. To support cognitive and physical affordance in the VR environment effectively, the design of the results and feedbacks should be realistic which is constituent of sensory affordance. In order to strengthen cognitive and physical affordance, more sensory aspects need to be considered than visual elements currently [36].

3.3 Step 2: Identifying Affordance Features

In step 2, affordance features are identified. Affordance feature is the components which should be considered to design affordance. The-stage-of-action model in step 1 is used to examine users' perspectives. Each stage in the model provides the ideas of users' specific actions and therefore the important features of objects, backgrounds, or agents in the VR game to display.

As seen in Table 1, the affordance features identified from previous research are insufficient for VR games. Minocha [31]'s affordance features fit to the learning in the VR environment, which did not cover the characteristics of the game. Bower [5] is missing the VR element by defining the elements of the overall media for the e-learning. Moreover previous studies did not consider the viewpoint of enhancing social presence.

Table 1. Affordance features

	Bower(2008)	Minocha et al(2017)	This rsearch(2018)
Environ- ment	e-learning	VR learning (Google Expedition)	HMD VR game
<i>Features</i>	<i>Media affordance</i>	<i>360-degree visual authenticity</i>	<i>First-person view</i>
	<i>Spatial affordance</i>	<i>360-degree navigation</i>	<i>3D journey</i>
	<i>Temporal affordance</i>	<i>3D view</i>	<i>Visual authenticity</i>
	<i>Navigation affordance</i>	<i>Emphasis</i>	<i>Stereoscopic 3D image</i>
	<i>Emphasis affordance</i>	<i>First-person perspective</i>	<i>Goal-oriented actions</i>
	<i>Synthesis affordance</i>	<i>In-situ contextual information</i>	<i>Emphasis</i>
	<i>Access-control affordance</i>	<i>Simulations</i>	<i>Action-based response</i>
	<i>Technical affordance</i>	<i>Single-user handling</i>	<i>Unintentional response</i>
	<i>Usability</i>	<i>Synthesis</i>	<i>Social environment</i>
	<i>Aesthetics</i>	<i>Visualization</i>	
	<i>Reliability</i>		

By describing the process by action from the users’ perspectives with the stage-of-action model, following critical affordance features of VR game are driven:

- (1) First person view: the VR game takes first person narrative. Therefore the background images and all the elements should be displayed along with the player’s movement.
- (2) 3D journey: Users are usually moving around to achieve a goal instead of staying at one place.
- (3) Visual authenticity (3D): Background image should be realistic as much as possible. Authentic visualization such as high-fidelity pictures and 360-degree physical view of the space helps users to sense and experience the virtual space which hard to visit in real life.
- (4) Stereoscopic 3D image: Things in the VR should be stereoscopic so that it looks real. This realism is about accurate representation of objects, events and people.
- (5) Goal-oriented actions: In the VR game, a player move his/her head, arms, legs, and body to achieve goals. Goal-oriented actions should be reliably and quickly designed as the same as the player expect.
- (6) Emphasis: Emphasis is some signs or things to highlight important information in the view. This additional information includes visually or audibly emphasized ways. It increases arousal level and cognitive spatial ability.
- (7) Action-based response: The VR system should be designed to respond to the actions of player properly and quickly.
- (8) Unintentional response: Except action-based response, VR environmental interaction depending on users’ action such as looking around and touching things which are not directly related with game goals. This is more related interactivity to feel presence in VR environment.

- (9) Social response: The VR environment is social-interaction friendly so that the player can be with other beings, play the same game with other players networked or interact with things in the VR.

3.4 Step 3: Two-Dimension Framework for Affordance Features of VR Game to Enhance Presence

In order to ensure all important affordance features are identified, we should consider what kinds of point of view we use to identify affordance features. Based on the previous literature review, two-dimension framework is induced. Axes of the framework are *Media Perspective* and *Interactivity Perspective*. Media Perspective is the key features which VR game should include: *VR Features vs. Game Features*. VR Features is related with the characteristics of VR media which is distinguish from other environment such as PC or mobile while Game Features are related with various game dependent factors. Interactive perspective is a key factor in VR game to support presence: *Interactive Control* and *Interactive Response* [42]. Among many other dimensions and features of interactivity, the two very basic components are used in this study. All the features should be implemented to enhance spatial presence or social presence of virtual world in VR.

The affordance features identified are categorized into each cell so that we can make sure the affordance features are identified enough to cover for every important aspect (See Table 2).

Table 2. Affordance feature categories based on media and interactivity perspectives

		Media	
		Game features	VR features
Interactivity	Interactive control	<ul style="list-style-type: none"> • Goal-oriented action • Emphasis 	<ul style="list-style-type: none"> • First-person view • 3D journey • Visual authenticity (3D) • Stereoscopic 3D image
	Interactive response	<ul style="list-style-type: none"> • Action-based response • Social response 	<ul style="list-style-type: none"> • Unintentional response

4 Discussion

This study clarified the concepts of presence, interactivity, and affordance which are critical to maximize user experience in VR game. Based on the conceptualization, we developed the three-step IAF model, which includes (1) a stage-of-the-action model for VR game environment that defines behaviors in the VR space from the user’s point of view, (2) affordance features identified, and (3) a two-dimension framework to ensure that affordance features for VR games to cover all perspectives. Affordance features are required to enhance the VR effects of the action for each stage of the model. Two perspectives of the matrix are (1) VR game dimension which includes VR focused aspect and game focused aspect and (2) interactivity dimension which are control aspect and response aspect.

The results of our study will be able to bring a comprehensive and consistent academic approach towards the roles of affordance and presence, which strengthen user experiences in VR games. This approach will be developed further with the changes of detailed elements in compliance with the advance of VR technologies. In practice, the IAF framework will be useful to identify the key elements of user-oriented design and services for the VR market which has not been expanded as expected.

Perceived presence can be different based on person's attributes, interests, and attention to the mediated spatial environment and arousal level [2]. Wirth et al. [41] also noted that cognitive-spatial abilities of users influence perceptions of spatial presence. This study, however, only focused on the characteristics of VR, not on personal factors. As well, there are many different environments to experience VR games, but this study investigates HMD-based VR environment, which currently are most widely used. Further studies on various VR environments could be done based on the results of our study. Finally, empirical studies will be needed to hypothesize and verify the relationships between the concepts driven from the framework and stage-of-the action model in this study.

At this very moment, the VR technologies are advanced rapidly and therefore the concepts and the framework proposed in this study will shed light on exploring the VR technologies in diverse disciplines.

References

1. Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M.: Equilibrium theory revisited: mutual gaze and personal space in virtual environments. *Presence Teleoper. Virtual Environ.* **10**(6), 583–598 (2001)
2. Baumgartner, T., Valco, L., Esslen, M., Jancke, L.: Neural correlate of spatial presence in an arousing and non-interactive virtual reality: an EEG and psychophysiology study. *Cyberpsychol. Behav.* **9**, 30–45 (2006)
3. Bicocca, F.: The Cyborg's dilemma: progressive embodiment in virtual environments. *J. Comput. Mediat. Commun.* **3**(2) (1997). <https://doi.org/10.1111/j.1083-6101.1997.tb00070.x>
4. Biocca, F., Harms, C., Burgoon, J.: Toward a more robust theory and measure of social presence: review and suggested criteria. *Presence Teleoper. Virtual Environ.* **12**(5), 456–480 (2003)
5. Bower, M.: Affordance analysis – matching learning tasks with learning technologies. *Educ. Media Int.* **45**(1), 3–15 (2008)
6. Cardona-Rivera, R., Young, R.: A cognitivist theory of affordances for games. In: *Proceedings of the Digital Games Research Conference: DeFragging Game Studies (DiGRA 2013)*, Atlanta, GA, USA (2013)
7. Coyle, J.R., Thorson, E.T.: The Effects of progressive levels of interactivity and vividness in web marketing sites. *J. Advert.* **30**(3), 65–77 (2001)
8. Digi-Capital. <http://digi-capital.com>
9. Deterding, S., Sicart, M., Nacke, L., O'Hara, K., Dixon, D.: Gamification. using game-design elements in non-gaming contexts. In: *Proceedings of Human Factors in Computing Systems Conference*, Vancouver, BC, Canada, pp. 2425–2428 (2011)
10. Fox, J., McEwan, B.: Distinguishing technologies for social interaction: the perceived social affordances of communication channels scale. *Commun. Monogr.* **84**(3), 298–318 (2017)

11. Gaver, W.: Technology affordances. In: Robertson, S.P., Olson, G.M., Olson, J.S. (eds.) *Proceedings of the ACM CHI 91 Human Factors in Computing Systems Conference* 28 April–5 June, New Orleans, Louisiana, pp. 79–84 (1991)
12. Gibson, J.: *The Ecological Approach to Visual Perception*. Houghton Mifflin Co., Boston (1979)
13. Grabarczyk, P., Pokropski, M.: Perception of affordances and experience of presence in virtual reality. *AVANT* **VII**(2), 25–44 (2016)
14. Hartson, H.: Cognitive, physical, sensory, and functional affordances in interaction design. *Behav. Inf. Technol.* **22**(5), 315–338 (2003)
15. Heeter, C.: Implications of new interactive technologies for conceptualizing communication. In: Salvaggio, J.L., Bryant, J. (eds.) *Media Use in the Information Age: Emerging Patterns of adoption and Computer Use*, pp. 217–235. Lawrence Erlbaum Associates, Hillsdale (1989)
16. Hutchby, I., Barnett, S.: Aspects of the sequential organization of mobile phone conversation. *Discourse Stud.* **7**, 147–171 (2005)
17. Klimmt, C., Hartmann, T., Frey, A.: Effectance and control as determinants of video game enjoyment. *CyberPsychol. Behav.* **10**, 845–847 (2007)
18. Klimmt, C., Vorderer, P.: Media psychology “is not yet there”: introducing theories on media entertainment to the presence debate. *Presence Teleoper. Virtual Environ.* **12**, 346–359 (2003)
19. Klimmt, C., Vorderer, P.: Interactive media. In: Arnett, J.J. (ed.) *Encyclopedia of Children, Adolescents, and the Media*, pp. 417–419. Sage, London (2006)
20. Lee, J.: *Internet and online game*. Communication Books (2001)
21. Lee, K.M.: Presence, explicated. *Commun. Theory* **14**, 27–50 (2004)
22. Lee, S., Shin, J.: Analysis of a Methodology of Design of Mobile Game Contents Based on Users’ Behavioral Patterns - Based on the Affordance Theory. *Preview Korean J. Digit. Mov. Image* **9**, 93–117 (2012)
23. Liu, Y.: Developing a Scale to measure the interactivity of websites. *J. Advert. Res.* **43**(3), 207–216 (2003)
24. Lombard, M., Ditton, T.B.: At the heart of it all: the concept of presence. *J. Comput. Mediat. Commun.* **3**(2), 1083–6101 (1997)
25. Majchrzak, A., Farai, S., Kane, G., Azad, B.: The contradictory influence of social media affordances on online knowledge sharing. *J. Comput. Mediat. Commun.* **19**, 38–55 (2013)
26. Mateas, M.: A preliminary poetics for interactive drama and games. *Digit. Creativity* **12**(3), 140–152 (2001)
27. McComas, J., Pivik, J., Laflamme, M.: Children’s transfer of spatial learning from virtual reality to real environments. *Cyberpsychol. Behav.* **1**, 115–122 (1998)
28. McGrenere, J., Ho, W.: Affordances: clarifying and evolving a concept. In: *Proceedings of the Graphics Interface 2000*, pp. 179–186. Canadian Human-Computer Communications Society, Toronto (2000)
29. McMillan, S.J.: What is interactivity and what does it do? Paper read at Association of Education in Journalism and Mass Communication Conference, August, Phoenix, AZ (2000)
30. McMillan, S.J., Hwang, J.-S.: Measures of perceived interactivity: an exploration of the role of direction of communication, user control, and time in shaping perceptions of interactivity. *J. Advert.* **31**(3), 29–41 (2002)
31. Minocha, S., Tudor, A., Tilling, S.: Affordance of mobile virtual reality and their role in learning and teaching. *Proc. Brit. HCI* **2017**, 1–10 (2017)
32. Newhagen, J.E., Corders, J.W., Levy, M.R.: *Nightly@nbc.com*: audience scope and the perception of interactivity in viewer mail on the internet. *J. Commun.* **45**(3), 164–175 (1995)
33. Norman, D.A.: *The Psychology of Everyday Things*. Basic Books, New York (1988)

34. Norman, D.A.: *The Design of Everyday Things*. Doubleday, New York (1990)
35. Regia-Corte, T., Marchal, M., Cirio, G., Lécuyer, A.: Perceiving affordances in virtual reality: influence of person and environmental properties in perception of standing on virtual grounds. *Virtual Reality* **17**(1), 17–28 (2013)
36. Reid, D.: A model of playfulness and flow in virtual reality interactions. *Presence Teleoper. Virtual Environ.* **13**(4), 451–462 (2004)
37. Steuer, J.: Defining virtual reality: dimensions determining telepresence. *J. Commun.* **42**, 73–93 (1992)
38. Theodoulou, P., Avraamidou, L., Vrasidas, C.: Flow and the pedagogical affordances of computer games: a case study. *Educ. Media Int.* **52**(4), 328–339 (2015)
39. VR focus. <http://vr-focus.com>
40. Weber, R., Bates, C., Behr, K.M.: Developing a metric of interactivity in video games. In: Annual convention of the National Communication Association, San Francisco, CA (2010)
41. Wirth, W., Hartmann, T., Böcking, S., Vorderer, P., Klimmt, C., Schramm, H., Saari, T., Laarni, J., Ravaja, N., Gouveia, F.R., Biocca, F., Sacau, A., Jäncke, L., Baumgartner, T., Jäncke, P.: A process model of the formation of spatial presence experiences. *Media Psychol.* **9**(3), 493–525 (2007)
42. Wu, G.: perceived interactivity and attitude toward website. In: Roberts, M.S. (ed.) *Proceedings of the American Academy of Advertising*, pp. 254–262. University of Florida, Gainesville (1999)



Augmented Reality and Telestrated Surgical Support for Point of Injury Combat Casualty Care: A Feasibility Study

Geoffrey T. Miller^{1,2(✉)}, Tyler Harris³, Y. Sammy Choi³,
Stephen M. DeLellis⁴, Kenneth Nelson³, and J. Harvey Magee¹

¹ Telemedicine and Advanced Technology Research Center, United States Army
Medical Research and Materiel Command, Fort Detrick, MD 21702, USA

geoffrey.t.miller4.civ@mail.mil

² Eastern Virginia Medical School, Norfolk, VA 23501, USA

³ Womack Army Medical Center, Fort Bragg, NC 28310, USA

⁴ United States Army Special Operations Command,
Fort Bragg, NC 28310, USA

Abstract. Providing surgical care in remote environments presents a significant challenge. Telepresence and telesurgery have the potential to bridge the gap between definitive care and nonsurgical critical care for prolonged field care scenarios. This feasibility study investigated several key questions regarding the suitability of these technologies for this application. First, what are the technology requirements and minimum specifications for telestration capabilities between a surgical specialist at a Medical Treatment Facility and a remote non-surgeon in a far-forward environment using existing telecommunication systems within the US Army? Second, what training requirements are needed to prepare surgeons and non-surgeons to control lower extremity junctional hemorrhage, and to use associated telestration hardware, software and communications systems? Third, what is the transferability of this training paradigm and technology suite to a wider range of medical care and clinical procedural skills to anticipated future military medical care needs and environments? Our initial feasibility study indicates that telementoring and telestration using augmented reality (AR) systems appears well suited to providing surgical support and training across dispersed groups of medical providers. Forward surgical support using AR and telestration technologies are viable for point of injury surgical support and may be essential to filling this “missing middle” in the Combat Casualty Care continuum. We anticipate that the life and limb saving capabilities supported by this approach will be necessary in future Multi-Domain Battlefield Concept and in cases of remote and dispersed operations.

Keywords: Augmented reality · Telestration · Telementoring
Modeling · Simulation · Combat casualty care

1 Introduction

Providing surgical care in remote environments presents a significant challenge. Dispersed medical operations and Anti-Access/Area Denial environments place casualties at risk of extended delays to forward surgical care sites. Postponed surgical care exposes casualties to avoidable suffering, loss of function or even loss of life [1]. Delayed emergency surgical care increases complications to damaged tissue and infection rates [2]. Unnecessary complications increase the cost of providing care to wounded service members [3]. Far-forward telestrated surgery support promises to mitigate these risks by stabilizing combat trauma in situations where this care would otherwise not be available.

Military trauma treatment has shown remarkable improvement in survival rates from World War II through Operations Enduring and Iraqi Freedom, where “died of wounds” rates decreased from 19% to 9%. Much of the recent improvement in survival is attributable to the liberal use of tourniquets and to rapid evacuation from the point of injury to locations offering damage control surgery. Although tension pneumothorax and airway compromise are in the top three causes of preventable death on the battlefield, hemorrhage remains the top cause of preventable casualty death [4]. Current and anticipated military operating environments threaten access to the forward surgical care necessary for damage control surgery that controls this hemorrhage. The lack of sufficient forward surgical resources has emerged as a critical capability gap.

The current U.S. military situation involves lower numbers of troops dispersed over a vast operating area. Additionally, near peer military rivals threaten U.S. air supremacy and military overmatch to the point that Area Denial and Anti-Access environments are expected in future conflicts [5]. Currently, Special Operation Forces in Africa operate across such great distances that providing rapid access to a field surgical facility is impossible in many cases. These military factors lead to the anticipation that many future casualties will need aspects of their initial damage control interventions, including some surgical procedures and intensive resuscitation, performed in the field. Telemedicine and telepresence promise the ability to move subspecialty resuscitation far forward to field units that are too separated geographically or too dangerous due to enemy activity [6].

Telepresence and telesurgery have the potential to bridge the gap between definitive care and nonsurgical critical care for prolonged field care scenarios. AR telementoring for surgery has been demonstrated to be effective for craniotomy and for carotid endarterectomy on cadavers [7]. The first transatlantic telerobotic surgery was performed in 2002 by a surgeon in New York City who removed the gallbladder on a patient in France [8]. Remote telerobotic proctoring was shown to be effective for life and limb saving procedures on cadavers between specialty surgeons and residents [9]. We anticipate that the life and limb saving capabilities augmented by these technologies will be necessary in the approaching Area Denial and Anti-Access environments. U.S. Special Operations Forces are already requesting these capabilities for their operational needs.

For the reasons and rationale stated previously, our research team reviewed potential candidate systems to begin to bridge this surgical capability gap. A concept of

operations (see Fig. 1) was developed to provide a high-level view of the capability need and guide research and development of forward surgical support solutions. A demonstration study was conducted to evaluate candidate technologies, training requirements, and process model testing, leading to a focused research investigation into the effectiveness of AR and telestrated surgical support for point of injury combat casualty care.

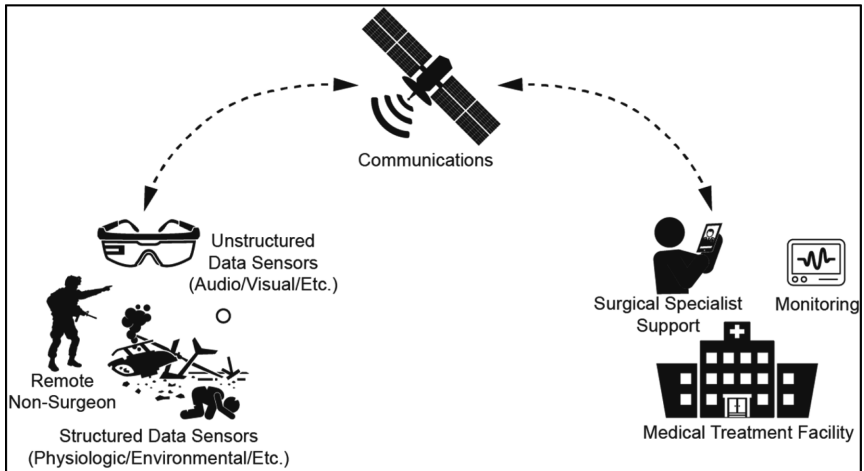


Fig. 1. Augmented reality forward surgical support concept of operations

2 Methods

An active-duty staff U.S. Army orthopaedic surgeon guided a U.S. Navy Physician Assistant (PA) through anterior exposure of the common femoral and external iliac arteries on a realistic surgical manikin using an AR wearable display. The surgeon was in a separate building, so he could not be seen or heard by the PA performing the procedure. The surgeon communicated with the PA using a Windows based personal computer with bidirectional voice communications and full motion video broadcast from the AR glasses. The PA used the AR glasses for all communication with the surgeon. The surgeon used the telestration capability of the glasses to superimpose instructions on the PA's visual field. These instructions appeared in the PA's surgical field, superimposed on the anatomy of the hyper-realistic surgical manikin. The telestrated instructions and voice communications were used to guide the PA through this damage control procedure.

Osterhout Design Group (ODG) R-7 Smartglasses (San Francisco, CA), a light-weight wearable on-visual-axis display, were furnished by BioMojo LLC (Cary, NC) with preloaded Librestream (Winnipeg, Manitoba, Canada) "Onsite Connect" telestration software and NuEyes® (Newport Beach, CA) magnification kits. The glasses were connected to a desktop workstation in a nearby building using an available wireless network. Surgical kits used were typical of what is available in a field surgery

setting. Forceps, Mayo scissors, Metzenbaum scissors, ring forceps, scalpels, Army-Navy retractors and Weitlaner retractors were available for the PA to use during the procedure. An Operative Experience Inc. (North East, MD) realistic anatomically correct manikin was used to simulate the anatomy and injury of a high femoral gunshot wound. A training classroom from the Fort Bragg Medical Simulation Training Center was mocked-up to represent an austere medical care environment. The PA and an additional nonsurgical assistant wore surgical gloves only for simulation purposes, as no infection control issues were presented by the manikin.

Prior to performing the procedure, the PA underwent crawl-walk-run pre-training sessions. The surgeon and PA received training with the ODG-R7 glasses (see Fig. 2). The training also contained a review of the indications, anatomy and technique for the procedure (see Fig. 3). These training sessions were performed to simulate actual training that would be conducted for selected telestration procedures if this technology were applied in a real-world situation. Other procedures selected could be fasciotomy and craniotomy.



Fig. 2. Technology training focusing on human interface with visualization and communication hardware and software. (Photo Credit: U.S. Army photo by Eve Meinhardt)

2.1 Feasibility Testing and Verification

Intensive, focused education and simulation-based training on anterior approach to the external iliac artery was provided. Verification of surgical skill competence was assessed on a newly developed surgical manikin. The PA and Surgeon were trained to perform the procedure while wearing the head-mounted display, to develop understanding and use of the augmented reality and telestration technologies. The team practiced the procedure, using the remote support technologies repetitively to assure

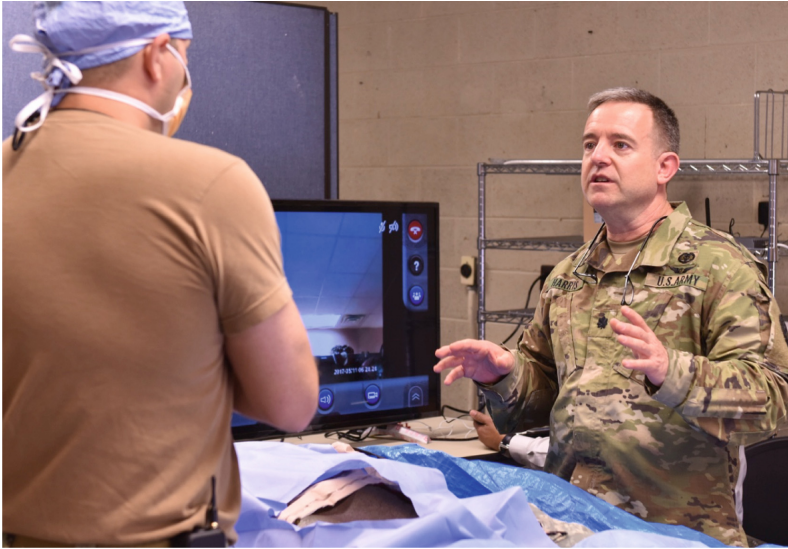


Fig. 3. Simulation-based procedural training between surgeon and non-surgeon. (Photo Credit: U.S. Army photo by Eve Meinhardt)

mastery of both the technology and procedure. Feasibility was assessed by a final demonstration and assessment in a simulated environment representing a remote, improvised surgical environment with a non-surgeon operator and a surgical specialist at a separate location.

2.2 Technical and Human Factors

Technical and human factors assessments will be conducted to evaluate the application, role and appropriateness of this surgical telestration and augmented reality capability. Assessments will be conducted to evaluate the usability of the system, workload and the comfort/confidence of the telestrating surgeon and the operating non-surgeon. Areas of assessment will include usability, ease of procedure, task load assessment, safety, efficiency, and operative time. Additional qualitative data will be collected through participant interviews and free-response instruments. Since this is an integration study we do not plan a control group.

Successful performance of this study and data analysis will guide submission of performance improvement strategies and follow-on studies that include field environment settings and a list of procedures considered by the United States Army Special Operations Command Deputy Chief of Staff-Surgeon. This list includes deep vessel access for control of junctional hemorrhage, abdominal cavity packing with temporary closure, upper extremity fasciotomies, lower extremity fasciotomies, advanced burn care resuscitation, burr hole craniotomy, resuscitative endovascular balloon occlusion of the aorta, external fixator placement, extremity vascular repair with shunt placement, and open fracture irrigation with debridement.

The feasibility study will focus on Navy Independent Duty Corpsman, Army Special Forces medics, Air Force Para-Rescue Jumpers and equivalents for the non-surgeon operators. These individuals are targeted for this study as they are prepared to provide medical services in austere environments, without ancillary, physician, logistic or medical evacuation support [10]. The surgical specialists will be recruited from military treatment facilities surgical specialties. Since this is a feasibility study we do not plan a control group. However, future studies will use control groups and additional procedures to deeply investigate telestrated surgery for study on human patients.

3 Results

Using AR, the surgeon remotely guided (see Fig. 4) the PA via telestration through anterior exposure of the femoral artery and external Iliac arteries (see Fig. 5). An additional non-surgeon held retractors, functioning as a surgical assistant. The surgical assistant only took direction from the PA and did not offer any guidance. The PA performed exposure and clamping of the proximal common femoral artery, which represented success for this procedure.



Fig. 4. Surgical specialist operating telestration console to guide anterior exposure of the femoral artery for bleeding control by a non-surgeon at a remote location. (Photo Credit: U.S. Army photo by Eve Meinhardt)

We successfully demonstrated a simulated, proof-of-concept use of lightweight wearable on-visual-axis display for surgical control. The PA/remote surgeon team, using augmented reality and telestration, were able to successfully perform junctional hemorrhage control, demonstrating the ability to project on-time, on-demand surgical expertise in a forward environment.



Fig. 5. U.S. Navy PA performing anterior exposure of the common femoral and external iliac arteries on a realistic surgical manikin using a wearable display with AR telestration connected to a remote surgical specialist. (Photo Credit: U.S. Army photo by Eve Meinhardt)

4 Discussion

Current military medical studies show there is a clear need for surgical capabilities within 60 min of injury in most deployed locations, regardless of how basic this capability may be [11]. Military medical specialists are in increasing need to perform damage control procedures to far forward environments when timely access to in-person surgical care is not possible due to distance or tactical considerations. Surgeons cannot be present in the far-forward area with each operational detachment team, increasing the need for advanced medical training and supportive telementoring and telestration technologies for specialized military medical care providers.

Telemedicine is defined as a set of medical practices without direct physician-patient interaction, via interactive audio-video communication channel employing tele-electronic devices [12]. Telestration, is defined as a technique for drawing freehand annotations over an image or video. Telestration has been found to be an essential functionality of telementoring systems [13]. An increasing body of knowledge is investigating the use of telementoring and telestration [12, 14, 15]. Telementoring has been reported in this literature as a natural fit in surgery demonstrating improved surgical practice, education, treatment and postoperative care [14]. While there has been much investigation into the technologies, cost effectiveness and safety of telementoring and telemedicine, there is little evidence regarding the rigorous study of clinical and educational outcomes [14]. The educational aspects of telestrated surgical support using augmented reality are a key focus of this study.

Telemedicine has advanced substantially from its initial use of telephone conversations to current advanced, real-time videoconferencing equipment, telementoring and

telestration making it well suited to supporting a means of transferring surgical knowledge across geographically dispersed individuals [15]. Advances in visualization and communication technologies offers the opportunity to train, equip and connect non-surgeon military medical providers with remote surgical experts to provide high quality emergency surgical care to far forward remote areas where access to evacuation to advanced surgical expertise may not be immediately available. However, clinical, user and mentor, technological and future research aspects of telementoring and telestration require further investigation and development.

In telestration-supported surgery, supervising surgeons draw or place virtual instruments on a device held over or remotely located to demonstrate procedures to assisted surgeons. The drawings or devices then appear superimposed on the surgical field, demonstrating procedural steps or important anatomy. Many current technologies present telestrated information off the central visual axis requiring the operative surgeon to look away from the operative field, compromising visual orientation for open procedures. Other devices held over the surgical field impair direct visualization of the field, degrading depth perception from loss of stereopsis. The size, weight, and requirement for fixed positioning further limits the field utility of these devices. We have demonstrated the use of a new lightweight wearable display with improved telestration capability to explore the feasibility of projecting surgical expertise forward in a field-able commercial-off-the-shelf device.

5 Future Directions

The next phases of our research and development of telestration for forward surgical support and telementoring will focus on evaluating the clinical benefits of these technologies on surgical interventions (including accuracy of clinical procedural skill performance as compared to telestrated targets, task completion and duration, and levels of mentoring and telestration required for procedure completion), educational benefits of the training programs and processes, quality of telestration and telementoring, and user satisfaction with both the remote telementoring and telestration aspects, and human-computer aspects. A mixed methods study has thus been designed to accomplish this purpose. This protocol was developed to evaluate the following aspects of providing surgical support to remote environments:

1. Technology requirements and minimum specifications for telestration capabilities between a surgical specialist at a Medical Treatment Facility and a remote non-surgeon in a far-forward environment using existing telecommunication systems within the US Army.
2. Training requirements to prepare surgeons and non-surgeons to control lower extremity junctional hemorrhage, and to use associated telestration hardware, software and communications systems.
3. Transferability of this training paradigm and technology suite to a wide range of medical care and clinical procedural skills to anticipated future military medical care needs and environments.

5.1 Training Protocol

A “Mastery Learning” model approach [16] will be employed to train and assess the surgical procedural skill performance of trainees (non-surgeons) and trainers (surgeons). Two surgical skills, anterior exposure of the femoral artery for bleeding control, and four-compartment fasciotomy, are the targeted procedures for the second phase of this project. The American College of Surgeons, Advanced Surgical Skills for Exposure in Trauma (ASSET) curriculum [17] will be used to teach the candidate surgical procedures.

The essential steps of this model include the following:

1. Establishment of a minimum passing mastery standard for each surgical procedural task, based on evidence-based or best practice standards,
2. Baseline assessment to determine appropriate level of difficulty of initial training activity needs,
3. Establishment of clear learning objectives, and performance indicators, sequenced as units ordered by increasing difficulty,
4. Engagement in training activities (e.g. skills practice, data interpretation) that are focused on reaching the objectives, and performance indicators,
5. Formative assessment and feedback to gauge surgical procedural task completion at the defined minimum mastery standard (e.g., repetitive error-free performance),
6. Advancement to the next surgical procedural training task when repeated measured performance meets or exceeds the standard, or
7. Continued practice or study on the surgical procedural training task until the mastery standard is reached.

This feasibility study will focus on Navy Independent Duty Corpsman, Army Special Forces medics, Air Force Para-Rescue Jumpers and equivalents for the non-surgeon operators. These individuals are targeted for this study as they are prepared to provide medical services in austere environments, without ancillary, physician, logistic or medical evacuation support [10]. The surgical specialists will be recruited from MTF surgical specialties. Since this is a feasibility study we do not plan a control group. However, future studies will use control groups and additional procedures to deeply investigate telestrated surgery for study on human patients.

Once surgeons and non-surgeons have completed the training component and achieved the mastery standard, they will advance to the surgical telestration environment and participate in technology training and simulation-based telestration scenarios focused on the surgical procedural training tasks. Simulation-based surgical telestration performance will be compared to the mastery learning training model and performance standard to evaluate the reliability (accuracy and consistency) of this model, as well as retention and transfer of training to the simulated environment.

6 Conclusions

Telementoring and telestration using AR systems appears well suited to providing surgical support and training across dispersed groups of medical providers. Forward surgical support using augmented reality and telestration technologies are viable for point of injury surgical support and may be essential to filling this “missing middle” in the Combat Casualty Care continuum. We anticipate that the life and limb saving capabilities supported by this approach will be necessary in future Multi-Domain Battlefield Concept and in cases of remote and dispersed operations. Continued rigorous investigation is needed to ensure safe and appropriate medical care in this environment as well as to inform the development and improvement of new and future technologies to support this capability.

Acknowledgements. Since the demonstration described in this report, the Army Medical Department’s Advanced Medical Technology Initiative Program has funded a research study for further investigation and development.

Disclaimer. The views expressed herein are those of the authors and do not necessarily reflect the official policy of the Department of Defense, Department of the Army, U.S. Army Medical Department or the U.S. Government.

Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government.



References

1. Pepe, P.E., Wyatt, C.H., Bickell, W.H., Bailey, M.L., Mattox, K.L.: The relationship between total prehospital time and outcome in hypotensive victims of penetrating injuries. *Ann. Emerg. Med.* **16**, 293–297 (1987)
2. Hull, P.D., Johnson, S.C., Stephen, D.J.G., Kreder, H.J., Jenkinson, R.J.: Delayed debridement of severe open fractures is associated with a higher rate of deep infection. *Bone Joint J.* **96-B** 379–384 (2014)
3. Thakore, R.V., Greenberg, S.E., Shi, H., et al.: Surgical site infection in orthopedic trauma: a case-control study evaluating risk factors and cost. *J. Clin. Orthop. Trauma.* **6**, 220–226 (2015)
4. Eastridge, B.J., Mabry, R.L., Sequin, P., et al.: Death on the battlefield (2001–2011): Implications for the future of combat casualty care. *J. Trauma Acute Care Surg.* **73**, S431–S437 (2012)
5. Gordon, J., Matusmura, J.: The Army’s role in overcoming anti-access and area denial challenges. No. W74V8H-06-C-001. RAND Arroyo Center, Santa Monica (2013)
6. DeSoucy, E., Shackelford, S., DuBose, J.J., et al.: Review of 54 cases of prolonged field care. *J. Spec. Oper. Med.* **17**, 121–129 (2017)
7. Shenai, M.B., Dillavou, M., Shum, C., et al.: Virtual interactive presence and augmented reality (VIPAR) for remote surgical assistance. *Neurosurgery.* **68**(1Suppl Operative) 200–207 (2011)
8. Marescaux, J., Leroy, J., Gagner, M., et al.: Transatlantic robot-assisted telesurgery. *Nature* **413**, 379–380 (2001)

9. Ereso, A.Q., Garcia, P., Tseng, E., et al.: Live transference of surgical subspecialty skills using telerobotic proctoring to remote general surgeons. *J. Am. Coll. Surg.* **211**, 400–411 (2010)
10. Rush, R.M.: Surgical Support for Low-Intensity Conflict, Limited Warfare and Special Operations. *Surg. Clin. North Am.* **86**, 727–752 (2006)
11. Kotwal, R.S., Howard, J.T., Orman, J.A., et al.: The effect of a golden hour policy on the morbidity and mortality of combat casualties. *JAMA Surg.* **15**, 15–24 (2016)
12. Budrionis, A., Bellika, J.G.: Telestration in mobile telementoring. In: eTELEMED 2013 The Fifth International Conference on eHealth, Telemedicine and Social Medicine, pp. 307–309 (2013)
13. European Commission. Guidelines on the qualification and classification of stand alone software used in healthcare within the regulatory framework of medical devices. http://ec.europa.eu/health/medicaldevices/files/meddev/2_1_6_ol_en.pdf
14. Augestad, K.M., Bellika, J.G., Budrionis, A., et al.: Surgical telementoring in knowledge translation—clinical outcomes and educational benefits: a comprehensive review. *Surg. Innov.* **20–3**, 273–281 (2012)
15. Miller, J.A., Kwon, D.S., Dkeidek, A., et al.: Safe introduction of a new surgical technique: remote telementoring for posterior retroperitoneoscopic adrenalectomy. *ANZ J. Surg.* **82**, 813–816 (2012)
16. McGaghie, W.C., Issenberg, S.B., Barsuk, J.H., Wayne, D.B.: A critical review of simulation-based mastery learning with translational outcomes. *Med. Educ.* **48**(4), 375–385 (2014)
17. American College of Surgeons Committee on Trauma. ASSET (Advanced Surgical Skills for Exposure in Trauma) Exposure Techniques When Time Matters. American College of Surgeons, Chicago (IL) (2010)



Cultivating Environmental Awareness: Modeling Air Quality Data via Augmented Reality Miniature Trees

Jane Prophet¹ , Yong Ming Kow² , and Mark Hurry³

¹ Goldsmiths College, University of London,
8 Lewisham Way, New Cross, London SE14 6NW, UK
j.prophet@gold.ac.uk

² City University, 18 Tat Hong Avenue, Kowloon Tong, Hong Kong
yongmkow@cityu.edu.hk

³ Independent Programmer, 165 Edmund Street, Beaconsfield,
Perth, WA 6162, Australia
mark@dwork.com

Abstract. The relationship between poor air quality and ill health concerns citizens and health practitioners the world over. An increasing number of Air Quality Data (AQD) apps quantify air quality numerically, yet many of us remain confused about the ‘real’ extent of air pollution, in part because we tend to ignore imperceptible pollution despite data alerting us to its danger. Further, even if we understand it, data about air quality alone does not sustain citizens’ interest in environmental issues. We report on our use of Augmented Reality (AR) to create an app that visualizes AQD as an affective miniature tree whose health corresponds to live AQD. We argue that AR is polyaesthetic, demanding a more embodied engagement from its users which deepens their understanding of AQD. Our participatory design study with 60 users shows the salience of using locally relevant imagery and working with users as co-designers to add features that support social interaction to design an app that gamifies citizens’ interactions with AQD.

Keywords: Augmented reality · Sustainability · Gamification

1 Introduction

Views from Hong Kong’s gleaming skyscrapers across the harbor can be breathtaking, but in January the wind direction changes, bringing the cool northeast monsoon which has become notable as much for the smog associated with it as for its increased rainfall. Pollutants from mainland China travel south blown in on these northeastern winds and our breath is taken away for all the wrong reasons as shown in Fig. 1. While January brings even more government warnings for children to play inside to avoid the worsening air, Hong Kong, like most cities around the world, has a year-long problem with pollution that it creates itself, predominantly as a result of emissions from diesel vehicles on its roads and waterways. The World Health Organization (WHO) found that “more than 80% of people living in urban areas that monitor air pollution are

exposed to air quality levels that exceed WHO limits.” This is not only in low-income cities. More than half (56%) of cities with 100,000+ citizens in high-income countries are experiencing pollution levels above WHO guidelines [1]. Despite the WHO data, many people are relatively unaware of the levels of pollution that they are living in as we tend not to consider poor air quality or check pollution levels until pollution becomes tangible in our daily lives. So how might a deeper awareness and an engagement with the science of air quality be achieved?

While sustainable HCI research has tended to focus on ‘improving’ public understanding of data (arguably an educational approach) and changing public behavior [2], a more dialogic process is gaining favor that situates science in what has been described as “circuits of culture” [3]. Instead of communicating a “correct” view of the science in question, the message is neutral. Proponents of the circuit model assert that meaning-making activities are dynamic, context-specific and change over time [3]. Sustainable HCI studies have used the term “discursive practice” [4] to describe a similar, situated, approach to design.

1.1 Awareness of Air Pollution is a Matter of Perception

To make healthier choices when planning our commutes and outdoor leisure activities in Hong Kong, we turned to computation, using air quality apps to get real-time AQD in numerical form. These apps (see Fig. 1) break AQD down into a number of categories that, together, present a complex picture. Even simplified green, amber and red ‘traffic light’ categorizations are accompanied by numeric readings, as seen in the Clean Air Network’s (CAN) app [5]. The HK Air Pollution app [6] displays overall air cleanliness levels according to different international guidelines. Often these app readouts did not correlate to what we saw or smelled. The air might look clear but data showed pollution levels were high, conversely sometimes the air smelled bad but,

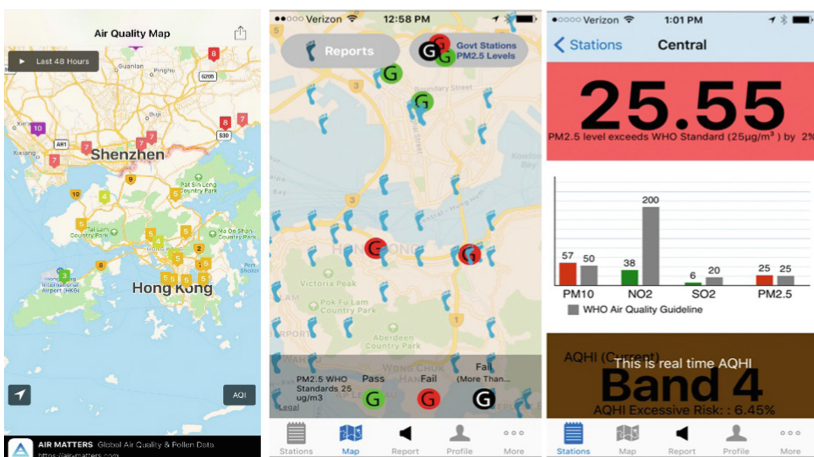


Fig. 1. Left: Air Matters app shows higher pollution in southern China than Hong Kong, January 2018. Middle: Clean Air Network app. Right: HK Air shows a breakdown of pollutants.

according to the apps, pollution was moderate. It turns out such mismatches between subjective sense of levels of pollution versus factual data about pollution, are typical.

Research shows [7] that our awareness of air pollution is a matter of perception, strongly influenced by what is observable and tangible, which does not always correlate with empirical measurements [8]. In addition, AQD is often broken down into the five key pollutants regulated under the Clean Air Act: ground-level ozone, particulate matter, carbon monoxide, sulfur dioxide, and nitrogen dioxide. Typically, AQD apps display information about each of these pollutants, creating a complex picture. Studies of public understanding of science show that the abstract yet complex nature of so-called ‘wicked problems’, like pollution and climate change, is notoriously difficult to understand [9], and that our lack of understanding is distancing [10]. Sustainable HCI has explored novel uses of personal informatics to gather data on how we behave in relation to these difficult-to-comprehend challenges [11, 12], to close this ‘distance’ [11, 13–15] and to prompt ‘transformative reflection’ [16].

Numerical data alone does not fully close the distance, so many projects visualize data to make it perceptible. Sustainable HCI research into invisible indoor air pollution found that visualization increased awareness and subsequent positive behavioral change [17]. But *how* such data is visualized, matters. There is somewhat surprising data about the use of supposedly affective icons, such as using polar bears to visualize climate change. Studies show that we are disinterested in images of places and animals [17] that are not part of our everyday lives [18] which applies to polar bears, for the majority of us that live outside polar bear habitat. The converse is also true, locally relevant images and cultural artefacts are more likely to carry scientific knowledge in ways we find engaging [19]. HCI studies have determined that an emotional engagement with data can be achieved by going beyond visualization, for example, by translating live data into multisensory experiences [13]. Therefore, we asked the question: *How can we create multisensory experiences to emotionally engage particular local users with scientific data about air quality?*

2 Designing to Improve Understanding of Air Quality Data

2.1 Dynamic Meaning Making by Citizens Through Networked Practices

In their study of risk analysis and mass media reporting of climate change, sociologists and media theorists propose a revised, more reflexive version of the older “circuits of culture model” which is relevant to our project. In this revised circuits of culture model “[m]eanings are remade in the contexts of social interaction at the local level as media texts are re-embedded in daily life” [3]. This helps to account for users making meanings about AQD and pollution from using our app and it places users at the center of meaning-making, recognizing their role in adapting and expanding the app and its features through dissemination and interaction with designers and researchers. We therefore took the circuit of culture model further, extending and formalizing the role of the user in meaning-making via our participatory design workshop which we discuss later.

Sociologists have argued that rather than being apolitical, young citizens’ “participation in social movements, rallies, protests, consumer boycotts all point to the possible displacement of traditional models of representative democracy as the dominant cultural form of engagement by alternative approaches increasingly characterized through networking practices” [20]. They go on to argue that young citizens’ attitudes are shaped more by their participation and interaction through social networks “which they themselves have had a significant part in constructing” [20]. Our previous research proposed that “mobile device practices in Hong Kong enable the significant extension of mobile leisure and gaming experiences into open and public spaces” [21]. Additionally, networking practices were central to the civic engagement and protests of Hong Kong’s Occupy Central movement [22, 23]. Pocket Penjing builds on these findings to create a mobile leisure experience that encourages civic engagement.

2.2 Local and Social Salience

Our decision to develop an AR app that only runs on smartphones was informed by the local relevance, or ubiquity, of mobile technologies. Statistics for smartphone use in Hong Kong [24] show that 85% of all Hong Kong residents over the age of 10 years have a smartphone. Of 15–49 year olds more than 97% have a smartphone and this drops by only 5% to 92% for 50–59 year olds.

In addition to choosing a locally-relevant platform, we drew on regionally-relevant imagery to represent AQD in our app, Pocket Penjing. The app takes its name from the Chinese “penjing”, tray plantings of miniature trees which pre-date bonsai. In contrast to Japanese bonsai, penjing place greater emphasis on landscape, environment and people’s relationship to the tree [25], reinforcing the idea that our virtual miniature trees are connected to the environment and to the human users that interact with them. We model a cherry blossom tree because of its ancient and ongoing iconic status in Chinese culture. Before we conducted our participatory design study, we produced a functioning prototype comprising a simplified computational simulation of a cherry tree that shows the impact of AQD on tree growth and displays the tree in AR.

2.3 Description of the Prototype App

The simulation described here is a prototype developed by our team of six, using sketches and group discussion to develop low fidelity prototypes, which we then iteratively adapted to make the stable prototype Android app which we describe here.

Scrape Live AQD, Create a Tree and Save It

The Android app is built using Unity3D. The introductory animation ends with an image of the canopy of a stylized cherry tree. By tapping on an empty area, users name their new tree. Up to three trees can be stored. Next, the user spins a 3D globe using a swipe motion, to choose one from a choice of flagged locations that represent live online air quality monitoring stations. The prototype uses data sources from live online air quality monitoring stations in Hong Kong [26] and Wuhan [27, 28]. The tree’s initial simplified environment is created from the selected location, using data recorded from the previous day in order to include actual pollution, rainfall and sunshine totals

over a 24-h period. Changes are input to the model via data from the selected station each time a saved tree is loaded and additional random inputs come from the users themselves as they seek to mitigate extreme weather or pollution by, for example, tapping relevant icons to add virtual shade to their tree to protect it from sun or wind, or to virtually water it (see Fig. 2).

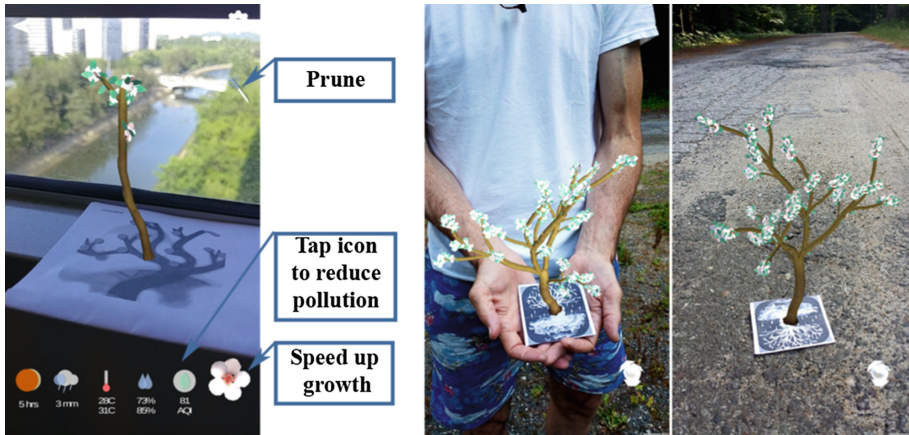


Fig. 2. Screensgrabs of the prototype. Left: AR tree aligned to final graphic marker placed on windowsill. Middle: Test marker design held in hands. Right: Test marker design on road. (Color figure online)

Display Lindenmeyer Tree in AR, Speed Up Growth and Prune the Tree

We use a basic deterministic 3D L-System [29, 30] to structure cherry blossom trees using turtle interpretation methods to calculate branching. The tree is defined at initialization, its final shape affected by three randomized branching angles (yaw, pitch, roll) that produce similar, but slightly differing, tree structures based on the same rules (see Fig. 2). The rules are informed by research into the growth and development of cherry trees [31, 32]. Pre-built sectional model components are assembled to form a three-dimensional tree. The app calculates the tree's daily growth amount from the selected live weather and AQD, incrementally scaling the tree's 3D components accordingly. The cumulative effects of applying each day's growth factor causes the tree to grow. We have not found any previous research that renders 3D forms, such as an L-Systems tree, in real-time from live data in AR.

A printed beer-mat sized reference image (see Fig. 2) registers the base of the tree's shoot/trunk and as soon as the tree has been made the AR window loads. When users first plant their tree only one set of historical data is available - the AQD that has just been scraped from the selected online air monitoring station. As this is the first data point, when users tap to plant their tree little or no growth will be visible. They have, in effect, planted a seed. On successive days, as people reload saved trees, users will see the tree growing, its rate and form based on the cumulative effect of the stored data. From when users first plant a seed, they have another option to fast forward tree

growth, by tapping the ‘speed up growth’ icon to see the shoot becoming a fully-grown tree in a couple of minutes. Figure 2 shows weather and air quality data displayed via small icons with numeric AQD readings underneath, icons turn red to alert users about data changes that will inhibit tree growth. As with real-life gardening, intervention is required to mitigate the weather data to enable the tree to grow despite, for example, overly polluted conditions. Users tap to alter temperature, pollution or rainfall. People can prune and shape the tree by removing branches (not the main trunk). There is no undo, just as when we cut the branch from a real penjing tree. After making the Android app prototype we conducted a participatory design workshop with 63 undergraduates to improve the prototype.

3 Using a Participatory Design Method to Develop the AR App

We conducted two, two-part formative participatory design workshops, which gave users one week to experience the app, reflect on its design, consider new design ideas, and share their ideas with the other participants and researchers. The format of workshops followed the method for conducting participatory design workshop studies, inspired by Gaver, et al.’s [33] notion of open design approaches, which are typically divided into three stages: (1) users are given an opportunity to get acquainted with the prototype through free-play or guided interactions; (2) users are asked to develop, categorize, and organize their own ideas into design concepts; and (3) the researchers will discuss these concepts with the users [33–35]. In the second part of the workshop, two researchers undertook a collaborative mind mapping exercise to capture emerging categories of ideas as they were described by users. We also audio and video recorded the participants’ sharing and discussion of their design concepts for further analysis. Following the workshops, we further discussed the ideas and implemented the most popular in the next version of the app.

3.1 Recruitment of Participants

We recruited 63 participants from an introductory course on computer games at our university in Hong Kong. The researchers obtained permission to conduct these workshops from the course instructor, who found value in using this activity as part of the students’ design assignments. The students of this course were divided into two classes, one running on Tuesday and the other running on Wednesday. In the Tuesday class, we recruited 33 participants, 23 (69.7%) were male and 10 (30.3%) female. In the Wednesday class, we recruited 30 participants, 17 male (56.7%) and 13 female (43.3%). Of all 63 participants, 63.5% were male and 36.5% were female. The gender breakdown is typical of university students who historically enrolled in this course. Participants were compensated with 5% towards their course grade. This course was conducted in English, the standard teaching language across all courses of the university.

3.2 Conducting the Two-Part Workshops

To accommodate the schedules of the two classes, we conducted two sets of the two-part workshop across two consecutive weeks - one for each class. In the first part, the researchers gave participating students of each class a one-hour briefing, including a live demonstration of the app. During the demonstration, two researchers helped students to download the app onto their Android phones and test it out immediately. The researchers also assisted any students who wanted extra information about, for example, the user interface. Students without an Android phone either shared use of the app with members of their group who had an Android device or used one of the five Android phones we provided for loan for the entire week. We only had two loan requests from students across the two classes. After the demonstration, students were given a standard ‘prompt’: “How do we redesign Pocket Penjing so as to most effectively persuade its players to care about environments in Hong Kong and Wuhan? Provide drawings, sketches, and other diagrams to illustrate your ideas.” Students continued working in pre-existing groups of four or five from their computer game classes. There were seven groups for Tuesday’s workshops and eight groups for Wednesday’s workshops. Each group was tasked with spending one week using the app, considering the prompt, and presenting their ideas for improving the app. They were asked to work together but to take the following approach: “To consider that every idea is a good idea, and not to remove any idea that they or their team-mates came up with.”

In the second section of the workshops, a week later, each group presented their ideas in 15 min to a total of three hours with the researchers acting as moderators. There had been no guidance or restrictions on how students might present their ideas. Every group made a group presentation using one Powerpoint, during which they all chose to give each member a chance to speak, each individual presenting their own ideas in turn, which was perhaps due to our instruction of “not to remove any idea,” but this turn-taking approach was also typical of the way students in that course had made previous presentations. Presentations were predominantly verbal with text-based slides, but most added illustrations or their own sketches. After each presentation, there was time for one or two questions, usually to elaborate on points covered in the presentation. In keeping with most other student presentations in this Hong Kong university, students were allowed to but did not ask their peers questions while the researchers did. Of the 15 total presentations, 13 groups gave us informed consent to audio and video record their presentations. We did not record the 2 groups that did not give consent. All students were assured that withholding consent would not be penalized in any way, such as by a reduced grade.

3.3 Data Analysis

To analyze any emerging categories of design concepts, and to map how such categories related to each other, we performed iterative coding during and after the participants’ presentations to capture [60] knowledge extracted during participatory design sessions. We used a collaborative mind mapping web tool, Coogler (coogler.it), shown in Fig. 3, that enables multiple users to construct a mind map together. Using Coogler, two

researchers made a mind map to document ideas, features, and designs that the students shared in each of the two workshop sessions. The two mind maps captured key points from all the presentations, including those made from groups that did not give consent to audio and video documentation. Both researchers accessed the same mind map document, each from their own laptops, and simultaneously added categories as the presentations took place. They could modify each other's categories and alter the mind maps live. No categories were discussed or developed by the researchers beforehand - all were determined, openly and inductively, in response to the presentations.

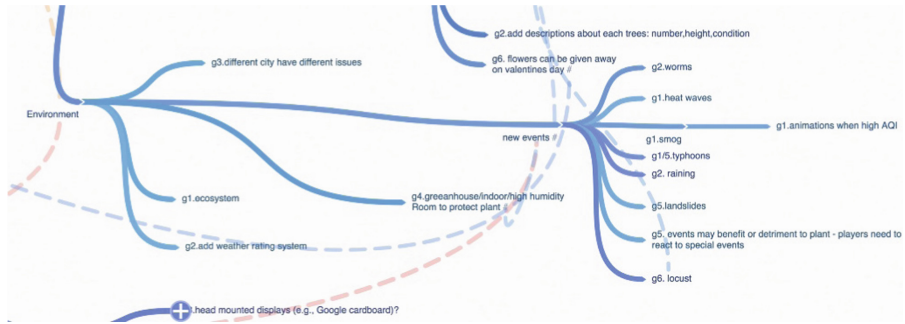


Fig. 3. Cropped section of Google mind map coding of one of the workshops.

After the workshops, we transcribed the audio recordings and performed more coding of the transcriptions modifying our initial codes. For example, during the mind mapping exercise, the two researchers had coded part of the social interaction features, such as flowers and seeds, into the category “virtual items.” Research team members’ iterative coding and analysis made it clear that certain virtual items are mediating social interactions - a salient theme found in many of the presentations. Thus, we conducted additional iterations of inductive and axial coding until themes presented within the findings section had emerged [36].

4 Findings

4.1 Preference for Locally Recognizable Features

Both the mind maps and transcripts of the videoed presentations show numerous references to the desire for additional features to the app that relate strongly to the local environment. Half the groups (6 out of 12 groups, 28 out of 63 participants), suggested the addition of regionally specific weather events and environmental conditions such as smog, typhoons, and heavy rain. Locally important, socio-cultural features were also suggested such this idea (see Fig. 4), “To encourage the players to interact with each other, we allow the players to plant conjoined trees (连理树). In Chinese culture it is similar to getting marr[ied to] the other players.” The participant referred to a particular phrase in a Tang Dynasty poem written by Bai Juyi (在天愿作比翼鸟,在地愿为连理枝), translated as “In heaven, let us



Fig. 4. Images from participatory design workshop presentations. Left: Conjoined trees as a symbol of love. Courtesy of Baidu Baike. Right: idea for sending trees to other users.

be birds with shared wings and body, and on earth, let us be trees conjoined.” The participant combines two Chinese cultural notions, penjing and a Chinese idiom about intimate social relationships. Another idea was to send trees to other players. Both ideas have social interaction at their core, conforming to suggestions made in early studies of social interaction in multiplayer games, to “reward the players who make these locations truly social environments” [18]. In the case of Pocket Penjing, the participants’ designs pointed to the salience of social interaction for understanding science.

4.2 Salience of Social Interaction for Understanding Science

Our participant designers suggested ways to integrate Pocket Penjing with social practices typical of the local Hong Kong student population. More than two thirds (49 out of 63) participants designed features to enhance social interaction which supports the theory that “[s]cience is a social process, yet scientists often pretend that it is not.” [10]. The most popular suggested new features were sharing images of the trees on Facebook, exchanging tips about how to cultivate a healthy tree and giving tree-related gifts, like seeds, produced by the tree, or sending someone an entire carefully cultivated and pruned tree. Making the tree part of established social festivals was also suggested, “[on] Valentine’s Day the players can give their flowers to their girlfriends or boyfriends and on [...] Mother’s Day they can [...] give the flower to their mothers.” (see Fig. 5). New features suggested by participants were sometimes coded as both game play and social interaction, for example, adding seeds and flowers that could be saved when users grew a healthy tree, was suggested by many participants. These items were described both as rewards for players, and as gifts that could be exchanged. There was a social dimension to the new difficulties, or levels, participants designed for the app. For example, the success associated with growing a tree in difficult situations was something not only personally motivating but also connected to a desire to show the results of overcoming that difficulty to their friends. This was as



Fig. 5. Images from participatory design workshop presentations. Two participants' ideas for multiple player social features and social media sharing.

much about sharing tips for how to cultivate a tree as it was about bragging. One participant saw the potential of the app to connect with users from beyond the local region and understand other climate and pollution issues, they suggested more options for scraping AQD from a wider range of “places around the world in regard to different climate differences from different places, besides just Hong Kong and Wuhan. [...] Perhaps different penjing got different climate conditions in favour in their growing. So that it will actually increase the interaction for the players and also make it more and more fun.”

Socializing and Learning Enhanced by More Difficult Individual Play Elements

Participants' came up with ideas to increase the number of simulations of local real-world phenomena, such as typhoons and smog, and often reasoned that this would result in better understanding of related pollution and weather. For example, this participant correlates exposure to risk to their tree with increasing awareness of local environmental issues, “in response to the large amount of rain during summer in Hong Kong, landslides may occur which may immensely affect the growth of the players' penjing. And then the players may be more careful in protecting the growth of the plant and may be informed by the environmental issues surrounding them.” One unexpected outcome was participants' desire to work harder to grow their tree, “introduce some sudden damages such as typhoon, bugs attacks or other disasters. So, when there is typhoon they have to use a barrier to protect the plant in order to stay away from damage. [...] players [...] pay more attention on the condition of the plant and then change their strategies”. This also introduces the idea of strategy, a game-like element that our prototype simulation did not support.

Most participants expected the app to be more game-like, with tokens and levels, and less a simple simulation. Although we had not described it as a game, we had described ‘playing’ with it in our prompt during the first workshop and our participant designers were from a games course, leading to a potential bias towards seeing the app as a game. However, sustainable HCI researchers investigating green transportation behaviors [15] using the UbiGreen Transportation Display mobile tool also found that, while they did not describe their product as a game, many users interpreted it as one and expected more game-like features such as points and levels.

4.3 The Importance of Real-World Interaction in AR Experiences

More than half (7 out of 12) of the groups of participants suggested additional real-world interaction to integrate the tree more fully into their local environment in ways that would involve users more bodily. 36 out of 63 presenters suggested more advanced AR and real-world interactive functions, including support of head-mounted displays, hand gesture recognition, audio and sunlight detection. Research has shown that AR prompts a deeper haptic or bodily engagement [37]. Participant designers echoed this with ideas for gesture recognition such as snipping their fingers (see Fig. 6) to prune the tree and modeling the impact of real-world physics on plant growth, typified by a sketch showing, “when we turn over the phone, the tree will grow upside down, and so a penjing may grow like this” (see Fig. 6).

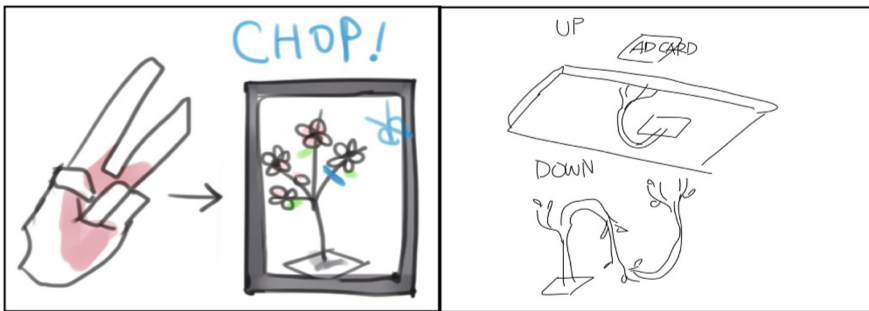


Fig. 6. Images from participatory design workshop presentations. Left: gesture recognition to prune the tree. Right: tree growing differently when phone is held upside down.

During the week between the two workshops, when they were exploring the app and coming up with ideas for the next version, some participants discovered our website [38] and were excited by screengrabs there that placed the AR tree in a variety of settings. They expanded on this idea, again making it more social, with the idea of placing a number of graphical markers together in parks so that multiple players could meet up to make a garden.

5 Discussion

5.1 Local (Poly)Aesthetics and the Salience of Social Networks for Learning

The aesthetic experience of Pocket Penjing reflects regional visual aesthetics of Chinese paintings of cherry blossom and Chinese paper cuts (see Fig. 7). However, as soon as the AR window loads it brings with it the aesthetics of the user’s offline world (see Fig. 7). While VR replaces visual cues from the local environment with an immersive artificial space, AR shows virtual objects within live camera images of our immediate surroundings, thereby going further than VR to “encourage us to occupy

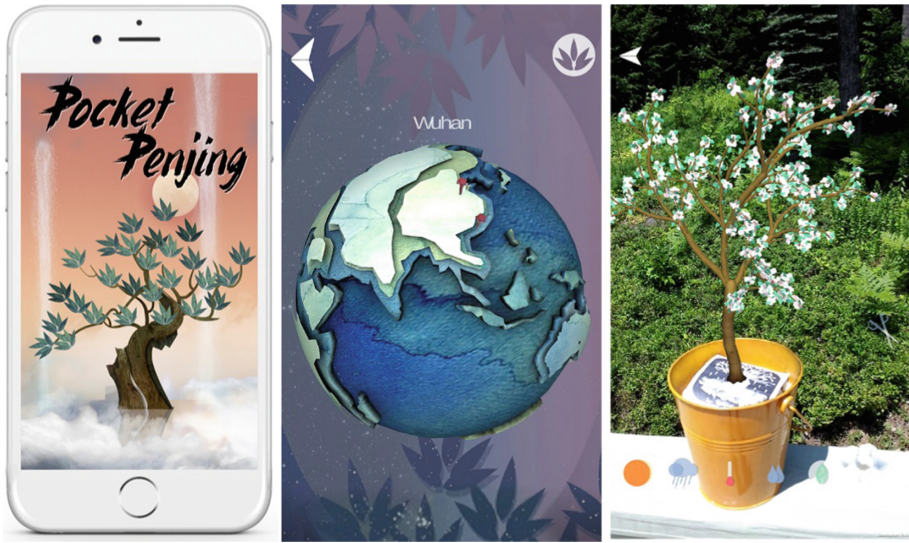


Fig. 7. Screenshots from the prototype. Left: opening image. Middle: papercut look to 3D pick globe. Right: AR cherry tree growing from marker placed in a tin bucket on a balcony.

two locations at once – to think of ourselves as both ‘here and there’ – situated in our immediate physical and cultural location as well as in the virtual world of the Internet” [39]. Our decision to use AR locates the virtual tree firmly in the local situation, emphasizing the importance of local context, “[o]ur augmented reality mobile apps can be seen as experiments through which we test hypotheses that mixed reality opens up the possibilities for more affective experiences” [39].

The way that AR blends the online and offline [40] means that Pocket Penjing’s visual aesthetics are not fixed, but rather they constantly emerge as different people use the app and take pictures of the trees, in different places. Our use of participatory design to co-develop features also multiply implications of the aesthetic. We therefore suggest that participatory design itself is polyaesthetic, bringing together multiple designers’ ideas, including features that take account of proprioception and embodiment. This brings us to third aspect of polyaesthetics in Pocket Penjing, that of multisensory involvement in AR [5]. For example, the use of a finger snipping action mentioned earlier, or another participant’s idea to shade the tree by cupping their hand over the phone screen to literally cast a shadow onto the screen. These gestures are typical of the way that our experience of using AR reminds us of our embodiment as we twist and turn our devices, and ourselves, to see virtual and real worlds combined on the screens and as we physically move in order to experience AR. AR affords a sense of spatial and emotional immersion that has been described as “polyaesthetic” [39] as it engages “multiple senses, and not only the senses of sight, hearing, and touch but proprioception as well” [37]. These multisensory and polyaesthetic affordances enhance learning [37], and support social interaction [41] and informal learning [42]. But what is the advantage of these complex and multilayered aesthetics? Do they motivate us to engage with AQD? Help us to learn? Or do they confuse and distract?

Constructivist learning theory, developed from psychology, extends the circuits of knowledge model [3], suggesting that we learn by constructing knowledge for ourselves, making meaning of the world both individually and socially [43, 44] and both these models are relevant when core users of the app are young citizens whose attitudes are shaped by networked social interaction [20]. Experts in learning technologies see synergy between AR and the power of situated and constructivist learning, arguing that “mobile devices equipped with AR experiences can enhance learning by situating data collection activities in a larger, meaningful context that connects to students’ activities at the real-world setting” [45]. Humans are social animals and these real-world settings are populated with other people. Playful engagement lends itself to social, even collaborative interaction with others. Therefore, we argue that a collaborative learning approach to science is likely to lead to higher achievement outcomes, and students will be more motivated and develop better social skills than learning science via more didactic approaches [37].

5.2 Persuasive Games, Learning and Sharing

The playful, rather than didactic, structure of our app has much in keeping with persuasive games [46] or ‘gamification’ [47]. It simulates the experience of having a small tree and learning what is necessary to nurture it as it responds, positively or negatively, to changing weather and air pollution. The values of our participant designers exist outside the game, and through their work with us they “focused on the social practices of playing the game, rather than the social practices represented in the game. [...] [games] are also media where cultural values themselves can be represented—for critique, satire, education, or commentary” [46]. This is part of what has been termed the ‘civic turn’ in HCI, the use of technology in civic action, engagement and participation in civic life [48]. The procedural rhetorical devices of league tables and scores, suggested as essential new features by our participants, are typical of many games and have been shown by sustainability HCI studies to enable comparisons between players and datasets [11] and incentivize engagement [49]. Kjledskov et al.’s HCI study of electricity consumption displayed on mobile devices showed comparative visualization to be beneficial to the increase of awareness necessary for the support of more sustainable behaviors [11]. While that study visualized different household’s energy usage, our mobile app supports multiple trees that can each represent AQD from a different monitoring station, hence enabling comparisons of the effect of air quality on trees in different geographic regions. Another game, Echo Chamber [50], uses procedural rhetoric, the practice of using computational processes persuasively, to make a game about the rhetoric and language of the debate the ‘wickedly complex’ topic of climate science. The designers found that the game was incidentally a learning tool for climate change, though its main focus was on utilising effective communication techniques. Like Echo Chamber, users of Pocket Penjing will learn about air pollution incidentally, while their main focus is on nurturing their tree and sharing that experience with others.

5.3 Scientific Data and Local Connections

By situating our participatory design workshops in Hong Kong, we have gone some way towards accounting for, and reflecting, those participants' shared cultural connections. Using techniques from game design, persuasive games “support existing social and cultural positions, but they can also disrupt and change those positions, leading to potentially significant long-term social change” [46].

Users interactions with data-driven rendered 3D cherry trees in Pocket Penjing create visual narratives “capable of providing clarity to the complicated and contested nature of toxic issues that would be considered controversial if stated in words” [52]. The narratives are not only visual though, as our sense of AQD is made tangible by being rendered as an AR experience that requires us to relate bodily to the data using multiple senses. AR experiences are not only multisensory, they are social, cultural and collaborative. The CAN app mentioned at the beginning of this paper (see Fig. 1) tip-toes towards more social engagement as it enables users to add comments which are then marked on a map with a footprint icon. However, there is no support of social interaction between commenters. Our prototype app provides numeric data about air quality but this is secondary to the representation of that data in a visual form. Like UbiGreen, our prototype was “not inherently social” [15] but, nevertheless, our participatory design study found that its graphics were conversation-starters, people wanted to know how the trees were progressing and to be able to share trees. Similar findings lead the UbiGreen team to plan future work to “explore the value of sharing application data among social groups” and we have now also begun to do that, as described below [63].

In summary, we have found it useful to converge changing trends that seem to run in parallel: the didactic *communication of science* of twenty years ago that has become the more circuit-based, or dialogic, *engagement with science* of more recent years; the move from didactic learning to constructivist and collaborative learning; the move from the “Modernist design in the 20th century [that] emphasized perfect integration of elements into a single unified form” [5] to the polyaesthetics of AR that exist in the aesthetic era of glitch and socially engaged art practice which emphasize the importance of locality and localness.

5.4 Implementing Ideas from the Participatory Design Workshops

Since completing the participatory design workshops we have implemented the following ideas: firstly, to increase the gamification and provide more visual feedback when users take actions. Participant designers wanted the addition of more international air quality index stations and we have added five, with more to follow (see Fig. 9). Our co-designers in the workshops suggested tapping icons was boring and more dynamic graphic representations were needed to show the tree's changes in health instead of it always looking perky. We have therefore developed a wider range of leaf and blossom textures to express health, for example the brown dry leaves in Fig. 8. Now, when the icons show a red alert, users access a toolbox and our first implementation of new visual tools to replace tapping on icons, in this case the low-water icon, is a watering

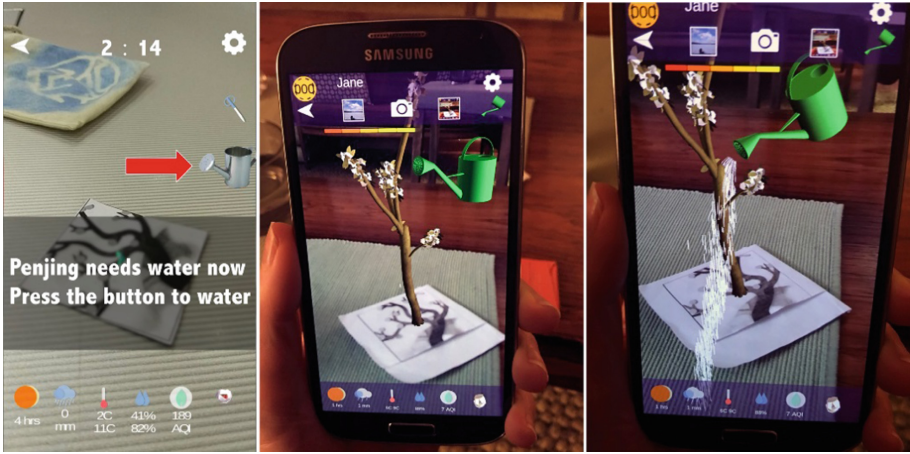


Fig. 8. Left: idea presented during participatory design workshop to replace tapping on icon to water tree, with image of a watering can. Middle: our implementation of the idea, plus implementation of idea to render leaves brown to show dryness. Right: our addition of animated tipping of can and particle systems water flow.

can animation that users activate (see Fig. 8). A number of different social sharing functions are in development but we have completed and implemented a gallery area within the app (see Fig. 9) and a related Facebook sharing feature (see Fig. 9).

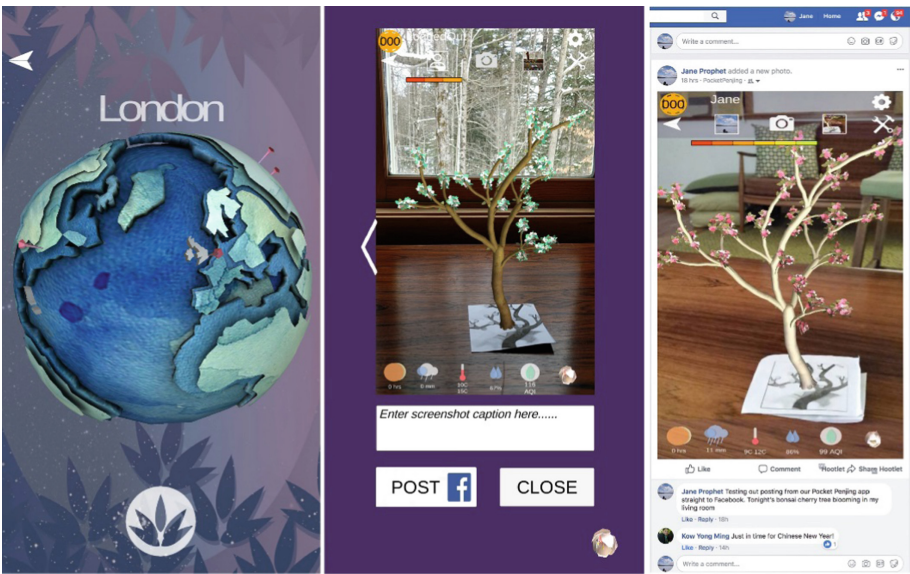


Fig. 9. Left: portion of the globe showing some of the additional AQI stations added. Middle: camera feature built into app allows users to create a gallery, scroll through their saved images and caption them for Facebook. Right: images are then immediately shared to Facebook.

5.5 Future Work

Designing with participants from Hong Kong in an English-speaking university setting presented few problems but we plan to conduct further studies in other localities and languages in order to explore the impact of language on the study and within the app itself. As described above we have begun to implement some of the features suggested from the participatory design workshops. We have more work to do in this regard, especially in relation to developing tokens, levels as part of a more gamified user experience. After the new features have been implemented we will conduct more naturalistic user studies that focus on how much more people understand about AQD after using the app in their daily life.

6 Conclusion

Pocket Penjing uses AR to enhance the sense of local context. The app supports informal, unstructured learning by making the focus of its persuasive simulation the tree and the users' relationship to it, while the scientific data that determines the tree's default condition is robust but downplayed. Instead attention is on our application of engaging data visualisation mechanics using real-time rendering of 3D trees. Learning how to engage with and interpret AQD is situated in a physical and cultural context in keeping with ideas that "knowledge is situated [53]. We designed the app on the premise that knowledge cannot be separated from the context in which it is learned [53] or, going further, that knowledge emerges through intra-actions between humans (scientists who develop metrics for measuring AQD, designers of App, users) and non-humans (ubiquitous computing networks, various mobile devices, air monitoring stations). Our use of participatory design studies shows that a more holistic approach to understanding AQD through embodied experiences situated in users' local real-world contexts, with playable media that have a number of features to support social interaction, are more powerful and engaging than abstract reasoning or relying on interpreting numeric AQD. Based on these findings, and on research studies showing that we engage and commit more if activities further social relationships, we argue that a deeper understanding of AQD may be fostered from embodied and socially-mediated interaction with data in local, blended offline/online spaces of AR.

Acknowledgement. The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 11608515). We also thank Pocket Penjing team members: Xinyi Chen, Graphic artist; Meichen Zhou, Research Associate; and Tsz Yan Andy Cheung, Research Associate.

References

1. Chriscaden, K., Osseiran, N.: Air pollution levels rising in many of the world's poorest cities (2016). <http://www.who.int/mediacentre/news/releases/2016/air-pollution-rising/en/>. Accessed 29 Jan 2018

2. Zhang, Z., Zhang, J.: A survey of public scientific literacy in China. *Public Underst. Sci.* **2** (1), 21–38 (1993)
3. Carvalho, A., Burgess, J.: Cultural circuits of climate change in U.K. broadsheet newspapers, 1985–2003. *Risk Anal.* **25**(6), 1457–1469 (2005)
4. DiSalvo, C., Sengers, P., Brynjarsdóttir, H.: Mapping the landscape of sustainable HCI. In: CHI 2010 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1975–1984 (2010)
5. Clean Air Network NGO. Clean Air Network (2009). <http://www.hongkongcan.org/>. Accessed 31 Jan 2018
6. The Air You Breathe. Hong Kong Air Pollution app [Internet] (2012)
7. Wakefield, E.L., Elliott, C.D., Eyles, J.D.: Environmental risk and (re) action: air quality, health, and civic involvement in an urban industrial neighbourhood. *Health Place* **7**(3), 163–177 (2001)
8. Barker, L.: Planning for environmental indices. In: Craik, K.H., Zube, E. (eds.) *Perceiving Environmental Quality*. ESRH, vol. 9, pp. 175–203. Springer, Boston (1976). https://doi.org/10.1007/978-1-4684-2865-0_10
9. Cohen, S., Demeritt, D., Robinson, J., Rothman, D.: Climate change and sustainable development: towards dialogue. *Glob. Environ. Change* **8**(4), 341–371 (1998)
10. Lorenzoni, I., Jones, M., Turnpenny, J.R.: Climate change, human genetics, and post-normality in the UK. *Futures* **39**(1), 65–82 (2007)
11. Kjeldskov, J., Skov, M., Paay, J., Pathmanathan, R.: Using mobile phones to support sustainability: a field study of residential electricity consumption. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2347–2356 (2012)
12. Pierce, J., Paulos, E.: A phenomenology of human-electricity relations. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2405–2408 (2011)
13. Jacobs, R., Benford, S., Selby, M., Golembewski, M., Price, G.G.: A conversation between trees: what data feels like in the forest. In: Proceedings of the SIGCHI Conference on Human Factors in Computing System, pp. 129–138 (2013)
14. Friedberg, E., Lank, E.: Learning from green designers: green design as discursive practice. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 1312–1323 (2016)
15. Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., Harrison, B.: UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1043–1052 (2009)
16. Ganglbauer, E., Fitzpatrick, G., Guldenpfennig, F.: Why and what did we throw out?: probing on reflection through the food waste diary. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 1105–1114 (2015)
17. Kim, S., Paulos, E.: In air: sharing indoor air quality measurements and visualizations. In: Proceedings of the Conference on Human Factors in Computing (CHI), pp. 1861–1870 (2010)
18. Webster, T., Dyball, M.: Independent Review of Beacons for Public Engagement Evaluation Findings. Final Report. RCUK, HEFCE and the Wellcome Trust (2010)
19. Csikszentmihalyi, M., Hermanson: What makes visitors want to learn? Intrinsic motivation in museums. *Museum News* **74**(3), 34–37 (1995)
20. Loader, B.D., Vromen, A., Xenon, A.: The networked young citizen: social media, political participation and civic engagement. *Inf. Commun. Soc.* **17**(2), 143–150 (2014)
21. Prophet, J.P.H.: Ubiquitous Alife in TechnoSphere 2.0. In: Ekman, B.J.D., Díaz, L., Sondergaard, M., Engberg, M. (eds.) *Ubiquitous Computing, Complexity and Culture*, pp. 254–266. Routledge (2015)

22. Mutsvairo, B., Harris, S.T.G.: Rethinking mobile media tactics in protests: a comparative case study of Hong Kong and Malawi. In: Wei, R. (ed.) *Mobile Media, Political Participation, and Civic Activism in Asia*. MCALIGI, pp. 215–231. Springer, Dordrecht (2016). https://doi.org/10.1007/978-94-024-0917-8_12
23. Kow, Y.M., Kou, Y., Semaan, B., Cheng, W.: Mediating the undercurrents: using social media to sustain a social movement. In: *CHI Conference on Human Factors in Computing Systems*, pp. 3883–3894 (2016)
24. Census and Statistics Department. *Information Technology usage: Women and Men in Hong Kong. Key Statistics*. Hong Kong: The Government of the Hong Kong Special Administrative Region. The Government of the Hong Kong Special Administrative Region, 27 July 2017
25. Ming, J., Liao, H., Chen, H., Wang, M.: Bonsai (penjing) systematic classification. *J. Nanjing Forestry Univ.* **25**(6), 59–63 (2001)
26. World Air Quality Index. *Hong Kong Air Pollution: Real-time Air Quality Index (AQI)* (2018). <http://www.aqicn.org/city/hongkong>. Accessed 16 Jan 2018
27. World Weather Online. *Wuhan air quality monitoring station*. <http://www.aqicn.org/city/wuhan>. Accessed 07 Aug 2017
28. World Air Quality Index. *Wuhan Air Pollution: Real-time Air Quality Index (AQI)*. <http://www.aqicn.org/city/wuhan>. Accessed 07 Aug 2017
29. Lindenmayer, A.: Mathematical models for cellular interactions in development 1. Filaments with one-sided inputs. *J. Theoret. Biol.* **18**(3), 280–299 (1968)
30. Rodkaew, Y., Chuai-Aree, S., Suchada, S., Chidchanok, L., Chongstitvatana, P.: Animating plant growth in L-system by parametric functional symbols. *Int. J. Intell. Syst.* **19**(1–2), 9–23 (2004)
31. Prusinkiewicz, P., Hammel, M., Hanan, J., Mech, R.: L-Systems: from the theory to visual models of plants. In: *Proceedings of the 2nd CSIRO Symposium on Computational Challenges in Life Sciences*, pp. 1–32 (1996)
32. James, P., Measham, P.F.: *Australian Cherry Production Guide*. Cherry Growers Australia Inc. (2011)
33. Gaver, B., Pacenti, P.E.: Design: cultural probes. *Interactions* **6**(1), 21–29 (1999)
34. Deneff, S., Ramirez, L., Dyrks, T., Schwartz, T., Al-Akkad, A.A.: Participatory design workshops to evaluate multimodal applications. In: *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, pp. 459–462 (2008)
35. Tohidi, M., Buxton, W., Baecker, R., Sellen, A.: User sketches: a quick, inexpensive, and effective way to elicit more reflective user feedback. In: *Proceedings of the 4th Nordic conference on Human-Computer Interaction: Changing Roles*, pp. 105–114
36. Corbin, J., Strauss, A.: *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage, Thousand Oaks (2007)
37. Bolter, J.D., Engberg, M., MacIntyre, B.: Media studies, mobile augmented reality, and interaction design. *Interactions* **20**(1), 36–45 (2013)
38. Prophet, J., Hurry, M., Kow, Y.M.: *Pocket Penjing Homepage* (2016). <http://www.pocketpenjing.com/>. Accessed 01 Feb 2018
39. Engberg, M., Bolter, J.D.: Cultural expression in augmented and mixed reality. *Convergence: Int. J. Res. New Media Technol.* **20**(1), 3–9 (2014)
40. Prophet, J., Pritchard, H., E Asian Ubicomp and ALife: Roaming and homing with technosphere 2.0 computational companions. In: *CHI 2015 Extended Abstracts: Between the Lines: Reevaluating the Online/Offline Binary* (2015)
41. Asai, K., Sugimoto, Y., Billingham, M.: Exhibition of lunar surface navigation system facilitating collaboration between children and parents in science museum. In: *ACM*, pp. 119–124 (2010)

42. Falk, J.H.: Free-choice environmental learning: framing the discussion. *Environ. Educ. Res.* **11**(3), 265–280 (2005)
43. Wadsworth, J.: *Piaget's Theory of Cognitive and Affective Development: Foundations of Constructivism*. Longman Publishing, White Plains (1996)
44. Hein, G.: Constructivist learning theory (1991). <https://www.exploratorium.edu/education/ifi/constructivist-learning>. Accessed 24 April 2017
45. Dede, C.: *Interweaving Assessments into Immersive Authentic Simulations: Design Strategies for Diagnostic and Instructional Insights*. Educational Testing Service, Princeton (2012)
46. Bogost, I.: *Persuasive Games: The Expressive Power of Videogames*. MIT Press, Cambridge (2007)
47. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining gamification. In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pp. 9–15 (2011)
48. Johnson, I.G., Vines, J., Taylor, N., Jenkins, E., Marshall, J.: Reflections on deploying distributed consultation technologies with community organisations. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2945–2957 (2016)
49. O'Neill, S., Nicholson-Cole, S.: Fear won't do it: promoting positive engagement with climate change through visual and iconic representations. *Sci. Commun.* **30**(3), 355–379 (2009)
50. Heath, C., Vom Lehn, D., Osborne, J.: Interaction and interactives: collaboration and participation with computer-based exhibits. *Public Underst. Sci.* **14**(1), 91–101 (2005)
51. Knowles, B., Blair, L., Coulton, P., Lochrie, M.: Rethinking plan A for sustainable HCI. In: *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3593–3596 (2014)
52. Peeples, J.: Toxic sublime: imaging contaminated landscapes. *Environ. Commun. J. Nat. Cult.* **5**(4), 373–392 (2011)
53. Geertz, C.: *Local Knowledge: Further Essays in Interpretative Anthropology*. Basic Books, New York (1983, 2000)
54. Aristodemou, E., Boganegra, L.M., Mottet, L., Pavlidis, D., Constantinou, A., Pain, C., Robins, A., ApSimon, H.: How tall buildings affect turbulent air flows and dispersion of pollution within a neighbourhood. *Environ. Pollut.* **233**, 782–796 (2018)
55. DiSalvo, T.L., Jenkins, T., Lukens, J., Kim, T.: Making public things: how HCI design can express matters of concern. In: *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2397–2406 (2014)
56. The European Travel Commission (ETC). ETC Country Reports: China (2014). http://www.tourism-generis.com/_res/file/3721/49/0/ETC_Country_Reports_China.pdf. Accessed 31 Jan 2018
57. Openweather. Weather forecast in Wuhan, China. <http://www.weatherforecastmap.com/china/wuhan/>. Accessed 16 Jan 2018



Enhancing Audience Engagement Through Immersive 360-Degree Videos: An Experimental Study

Ayoung Suh^(✉), Guan Wang, Wenying Gu, and Christian Wagner

School of Creative Media, City University of Hong Kong,
Kowloon Tong, China

{ahysuh, c.wagner}@cityu.edu.hk,

guanwang3-c@my.cityu.edu.hk, wenyingu@um.cityu.edu.hk

Abstract. The use of 360-degree videos to engage audiences in diverse contexts is increasing. While 360-degree videos have the potential to create new value in enhancing audiences' viewing experiences, they often decrease audience engagement by causing motion sickness in an immersive environment. Despite increasing scholarly and practical attention to the effect of 360-videos on audience engagement, the question of how to enhance it through immersive 360-degree videos remains unanswered. Therefore, this study empirically examined the effects of different display types and viewport dynamics on audience engagement using data collected from 60 subjects during a laboratory experiment. The results show that an audience's viewing experience in an immersive environment is influenced by the joint effects of display types and viewport dynamics. By explaining the mechanisms by which audiences are engaged with 360-degree videos, this study contributes to resolving previous inconsistent findings regarding the effect of immersive technology on audience engagement.

Keywords: 360-degree video · Audience engagement · Display device
Viewport dynamics · Presence · Motion sickness

1 Introduction

As a new branch of immersive technology, 360-degree videos are gaining popularity in many contexts, including education [5, 26], marketing [9], journalism [54], and clinical training [20]. The growing use of 360-degree videos is aimed at increasing audience engagement [2]. By providing a spherical view of the camera's surroundings, 360-degree videos allow the audience to control the camera's orientation and view content in an immersive environment [31]. Researchers have found that 360-degree videos enhance audience engagement by increasing the audience's sense of presence and immersion, which increases their perceived enjoyment while viewing the videos [48, 50]. However, evidence shows that watching a 360-degree video can also cause negative experiences, such as motion sickness, boredom, and distraction [2, 28, 44], which reportedly undermine audience engagement. To harness the benefits of

360-degree videos, it is important to understand how to enhance audience engagement by simultaneously increasing positive experiences and inhibiting negative experiences.

Research has suggested that different types of display devices influence audience engagement while viewing 360-degree videos. For example, it has been found that head-mounted displays (HMDs) with stereoscopic capabilities provide full immersion, which enhances audience engagement [12, 31, 57]. HMDs are also beneficial for navigational purposes because they provide a natural interaction mode, which allows audience members to control the camera's orientation by turning their heads to see the surrounding environment [31]. However, some researchers have claimed that this natural interaction might override the benefits of using HMDs. In addition, because HMDs enable viewers to change their orientations in seconds in a highly immersive condition, it can cause more motion sickness and visual discomfort than other display devices [2]. Van den Broeck et al. [2] argue that HMDs' shortcomings inhibit exploration and that some people prefer watching 360-degree videos on mobile devices due to the simplicity of exploration and their familiarity with the navigation controls. Mobile devices also have other advantages for users, such as the ability to quickly scan their surroundings and a greater freedom of mobility [31]. Notably, conflicting views regarding the effects of display devices on audience engagement have yet to be reconciled.

Furthermore, it has been argued that, while the viewport dynamics of a virtual environment can influence audience engagement by creating a sense of presence, they can also cause motion sickness [2]. According to motion parallax theory [11], viewport dynamics create a sense of depth (i.e., a sense of the distance of an object) and enhance the vividness and realism of a virtual environment, which both ultimately increase the audience's sense of presence [13]. However, viewport dynamics have been found to be a source of motion sickness, especially when a video scene contains rapid motions [53]. In the case of 360-degree videos, the effects of viewport dynamics with combined camera and object movement should be further discussed to better understand audience engagement. Previous studies have focused primarily on the effects of either display devices or viewport dynamics. However, these separate research streams do not inform one another, which has limited our comprehensive understanding of why 360-degree videos often fail to engage audiences. Relatively little effort has been made toward exploring how display devices and viewport dynamics jointly influence audience engagement with 360-degree videos. Therefore, the present study seeks to answer the following questions:

RQ1: How do display types influence audience engagement with 360-degree videos?

RQ2: How do viewport dynamics influence audience engagement with 360-degree videos?

To answer these questions, this study develops hypotheses regarding the influences of display devices and viewport dynamics on presence, motion sickness, and audience engagement. We conducted a between-subjects experiment with display devices and viewport dynamics as between-subjects factors. Our results show that both display devices and viewport dynamics could significantly influence audience engagement. This study extends our understanding of the underlying mechanisms of audience engagement with 360-degree videos in several ways. The study firstly complements

and extends the existing literature, which has yet to empirically and thoroughly test the positive and negative influences of different types of devices on audience engagement. The study also examines how viewport dynamics influence presence, motion sickness, and engagement. Finally, it provides useful insights that could help content creators and engineers better understand audience engagement.

2 Related Work

2.1 Audience Engagement

Audience engagement, as a quality of audience experience, refers to the audience's positive reactions to technologies [39, 40]. Audience engagement has increasingly been studied to foster a deeper understanding of audiences' experiences in viewing 360-degree videos. In this study, following O'Brien and Toms [38], we conceptualize audience engagement as the extent to which an audience achieves deep cognitive, affective, and behavioral involvement with 360-degree videos.

Researchers have found that 360-degree videos encourage audiences to actively engage with content because of their increased display fidelity, which is achieved through navigable views and stereoscopy [31, 33]. Scholars commonly argue that two key factors—sense of presence and motion sickness—determine audience engagement with 360-degree videos [37].

Presence. Presence is the extent to which a user feels that he or she is in a particular place, even while physically situated in another place [19]. A 360-degree video can enhance presence by displaying a spherical viewing area and creating a natural visual experience. It can also increase the authenticity and realism of the viewing experience via a wider field of view [21, 27, 45]. By enabling viewers to rotate their views to look anywhere around them, 360-degree videos create a first-person view of an event or situation, which enhances presence [2, 44, 49, 54].

People who have a greater sense of presence are more likely to feel engaged in immersive environments [13]. According to presence theory, a technology that supports high levels of user presence (i.e., a sense of being there) fosters users' motivation to engage with the technology by enabling focused and naturalistic interactions [23, 60]. Presence theory suggests that presence in an immersive environment is strongly correlated with positive emotions, such as enjoyment, playfulness, pleasure, and fun [8, 55, 59]. Motion parallax theory [11] suggests that a first-person perspective may lead viewers to feel that they have actually experienced the events or situations shown in the 360-degree video, which can lead to increased cognitive engagement [7, 48]. It has also been found that presence increases viewers' participation in activities (e.g., viewing, commenting, and positing in relation to videos). Accordingly, we propose the following hypothesis:

Hypothesis 1: Presence is positively related to audience engagement.

Motion Sickness. While 360-degree videos have been found to offer the audience an increased sense of presence, some researchers note that motion sickness that is

engendered by 360-degree videos may decrease audience engagement. Motion sickness includes physical symptoms, such as fatigue, dizziness, and blurred vision [25]. When orientation signals that are transmitted by the eyes and the vestibular organs do not match, people experience sensory conflict, which can lead to motion sickness [15]. In the case of 360-degree videos, the cause of motion sickness is the conflict between the visual motion information transmitted by the video and the viewer's real-world perceptions [22].

Motion sickness draws attention away from the virtual environment, which makes it difficult for viewers to focus their attention on video content and thus decreases involvement [61]. Motion sickness can also increase viewers' discomfort and reduce their enjoyment [27]. Prior research has found that people may discontinue watching 360-degree videos to avoid the negative experience of motion sickness [10]. Thus, we propose the following hypothesis:

Hypothesis 2: Motion sickness is negatively related to audience engagement.

2.2 Display Type

Various types of display devices, including HMDs, mobile devices, and personal computers, can be used to view 360-degree videos. Viewers using HMDs can control the camera orientation by turning their heads to see the surrounding environment. With mobile devices, viewers can watch 360-degree videos through interactive screens by either clicking and dragging content horizontally and/or vertically or through mobile device sensors by manually moving their devices [62].

These display devices represent a continuum of degrees of immersion, which refers to a technology's capability to simulate and surround a user with layers of sensory information [61]. Research has suggested that people in highly immersive conditions feel significantly higher levels of presence. In the context of 360-degree videos, HMDs provide the highest degree of immersion and produce the strongest presence [12, 30, 43, 58]. Immersive devices isolate viewers from their physical environments and deprive them of environmental sensations [61]. HMDs can effectively block information from the physical environment and provide more natural and intuitive interactions, which can engage viewers and provide the sensation of being inside the video scene [2, 47]. Thus, we propose the following hypotheses:

Hypothesis 3a: HMDs elicit greater degrees of presence than mobile devices.

Hypothesis 3b: Presence mediates the relationship between display type and audience engagement.

Previous studies suggest that different display devices can cause different levels of motion sickness in an immersive environment [24]. Users in high-immersion conditions have more severe symptoms [18, 36]. Specifically, motion sickness is common when using contemporary HMD systems [6, 36]. Users of HMDs in motion who engage in interactive environments in which their visual perceptions of motion may not be aligned with their physically perceived motion may exhibit some degree of motion sickness [32]. Motion sickness may also occur in cases of detectable lags between head movements and the recomputation and presentation of the visual HMD display [17].

Sensory conflict exists between the visual motions in 360-degree videos and the sense of physical motion. This conflict between real-world motion perceptions (e.g., head movements) and visual motions is more salient in the HMD condition. Although viewers can control the direction of viewing by rotating their heads, head movements, such as leaning forward, have no impact on the visual stimulus [34]. By contrast, viewers using mobile devices to view 360-degree videos can use their peripheral vision to get congruent motion cues from the video background (i.e., the real world). Thus, we propose the following hypotheses:

Hypothesis 4a: HMDs elicit greater degrees of motion sickness than mobile devices.

Hypothesis 4b: Motion sickness mediates the relationship between display type and audience engagement.

2.3 Viewport Dynamics

The viewport is the area of the 360-degree video frame that is displayed at a given time [41]. There are two main types of viewports for 360-degree videos: (1) moving viewports (MVPs), in which the camera moves and turns to keep the subject in view while recording, and (2) static viewports (SVPs), in which a camera is statically placed in the center of the scene during shooting [2, 13].

Viewport dynamics have a significant influence on how the audience explores and experiences a 360-degree video. Previous studies have found that 360-degree videos with MVPs offer a superior viewing experience and elicit higher audience engagement [2]. In addition, they support a faster and greater sense of depth and sharpness than videos with SVPs [42, 56]. MVPs also provide better motion parallax cues, which enhance the sense of a video object's depth and naturalness [13] because they provide a first-person perspective that allows the audience to experience events or situations in the same way in which the video producer did [7]. Finally, MVPs can trigger viewers to change direction and explore the mediated environment, thereby enhancing interactivity and sense of presence [2, 13]. Thus, we propose the following hypotheses:

Hypothesis 5a: MVPs elicit a greater degree of presence than SVPs do.

Hypothesis 5b: Presence mediates the relationship between viewport dynamics and audience engagement.

The audience may want to visually follow a moving object in a 360-degree video [28]. Viewers' ability to actively change their view orientation creates a challenge for the experience of 360-degree videos. Although high speeds can help viewers quickly focus on moving objects, they may also increase motion sickness [28]. Regarding exploration behaviors, it has been found that viewport dynamics (i.e., the degree and direction of head turning) have a substantial effect on the degree of motion sickness in immersive environments [51]. SVPs allow viewers to follow moving objects by turning around quickly, whereas MVPs enhance the synchronization between such objects' motions and viewers' physical motion. MVPs reduce the speed of physical motions to support orientation change because they synchronize viewers' motions with the motion of the camera when viewers choose to follow moving objects. Thus, we propose the following hypotheses:

Hypothesis 6a: MVPs elicit a lesser degree of motion sickness than SVPs.

Hypothesis 6b: Motion sickness mediates the relationship between viewport dynamics and audience engagement.

3 Methods

3.1 Design

To test our hypotheses, we conducted an experiment with display type and viewport dynamics as between-subjects factors. Participants were randomly assigned to two display conditions (HMD or mobile device) and asked to view two 360-degree videos that contained different viewport dynamics (moving and static). To control the influence of screen quality and screen size, we selected Google Nexus as our mobile device. Participants in the HMD condition used a standard cardboard viewer 2.0 that was equipped with the same Google Nexus phone. The viewport dynamics that were manipulated in this experiment had two types: SVP and MVP. The order of the two videos was fully counterbalanced across participants.

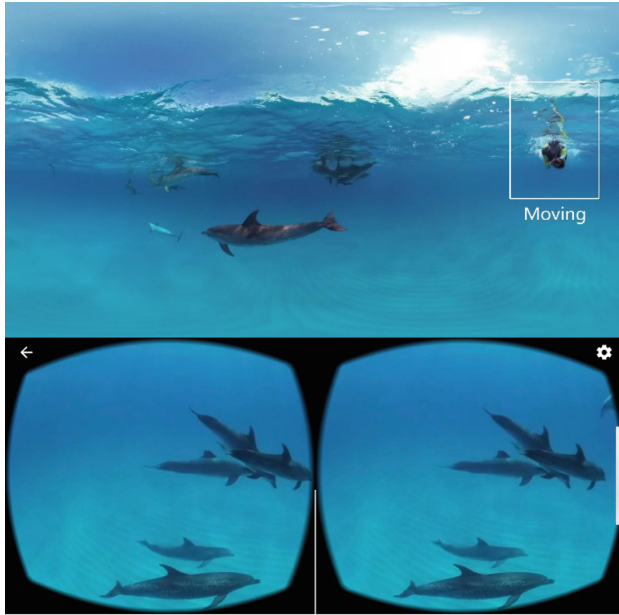
3.2 Stimuli

The video content in the experiment was fully controlled and counterbalanced. We first benchmarked the various existing examples of 360-degree videos and then evaluated the most suitable video content for this study. We selected the diving genre as our video stimulus because it is one of the most popular 360-degree video genres on YouTube, and diving videos are easy to control regarding the velocity of both the camera movement and the subject. We selected multiple 360-degree diving videos from YouTube and edited them into two four-minute videos per their viewport dynamics: MVP, in which the camera moved and turned to keep the subject in view, or SVP, in which the subjects moved within the scene yet the camera was static. Figure 1 shows screenshots of the two videos. Half of the observers viewed the MVP videos, followed by the SVP videos, and the other half watched the videos in reverse order.

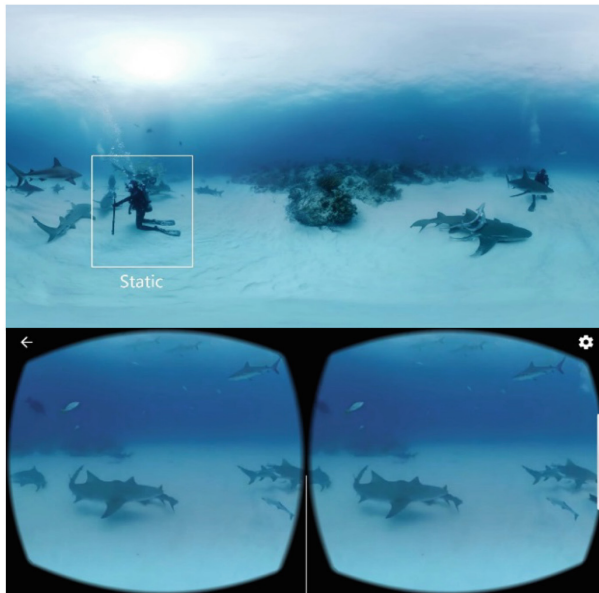
To enable a comparison of the differences in the videos' camera motions, all other factors were kept as equal as possible. For example, we controlled the features of the subject's motion and the distance from the camera to the subject(s) as much as possible. Table 1 describes the criteria applied to control all video content factors.

3.3 Participants

Our participants were recruited from a university in Hong Kong. We advertised for "A Fun 360-degree Video Study" on a university-facilitated online system that is used for recruiting experiment participants for various research purposes. The experiment lasted three days. A total of 64 participants both registered for and attended the study. We deleted responses that repeated answers. After invalid responses were removed, 60 of the 64 responses ($M_{age} = 22.9$, $SD_{age} = 4.49$, Female = 50%) were included for further analysis. Table 2 summarizes the respondents' demographic information.



a. Moving Viewport



b. Static Viewport

Fig. 1. Screenshots of the two videos used in the experiment

Table 1. Control criteria

Dimension	Control	Reason
Content genre	Diving	It is a popular genre; it is easy to control the motion velocity of the subjects and the camera
Duration	Approximately 3–4 min	It is the choice of most prior studies regarding the balance between interest and mental/physical load. A user may not view a 360-degree video for as long a duration as a regular video due to either feeling uneasy or experiencing motion sickness [1]
Visual quality	Same resolution rate (1080 s)	Visual quality is one of the most important factors that influence audience experience. Low visual quality can have a negative effect on the audience’s engagement and enjoyment of the videos, regardless of visual realism [29, 43]
Auditory quality	Same background music; same earphones	Sound plays a role in emotional reactions and presence [58]
Emotional intensity	Controlling the content and audio	Emotional content can increase presence [3, 12]

Table 2. Demographics of respondents

Item	Category	Frequency	Ratio
Gender	Male	30	50%
	Female	30	50%
Age	<21	18	30%
	21–25	24	40%
	26–30	15	25%
	>30	3	5%
Major	Science	4	6.7%
	Engineering	11	18.3%
	Humanities & Arts	17	28.3%
	Social Sciences	14	23.3%
	Business	12	20%
	Law	2	3.3%
Familiarity with 360-degree video	Strongly disagree	3	5%
	Disagree	11	18.3%
	Somewhat disagree	13	21.6%
	Neither agree nor disagree	9	15%
	Somewhat agree	17	28.3%
	Agree	5	8.3%
	Strongly agree	2	3.3%

3.4 Procedure

Because prior studies have suggested that people can feel discomfort in public environments [31], a maximum of five participants completed the experiment in the laboratory room at the same time. Upon arrival, participants were seated approximately two meters away from one another to ensure an unrestricted range of motion during viewing. Swivel chairs, which have been recommended for experiments involving 360-degree videos, were supplied [2, 14].

Each participant was randomly assigned to a single device condition. Specifically, 30 participants watched the 360-degree videos using an HMD (i.e., Google Cardboard), while 30 participants watched the videos using a smartphone. Participants were first introduced to the experiment devices to ensure familiarity, and a tutorial on the basic functions required to display 360-degree videos was offered. After the tutorial, the tasks in the experiment were explained to the participants. Both the tutorial and the tasks were reiterated verbally by the experimenter for each new round of videos. The participants were asked to answer a pre-test questionnaire containing questions related to demographic information and their familiarity with 360-degree videos. After the participants completed the pre-test questionnaire, the experimenter equipped the participants with headphones and asked them to watch the first video. Following the video, the participants completed a post-questionnaire containing subjective measures. Then, the participants watched the second video. After watching the second video, the participants were asked to select their favorite video and to answer an open-ended question by stating the reason(s) for their choice. Upon completion of the study, each participant was offered a coupon worth HK\$50 as an incentive for participating.

3.5 Self-report Measures

Measurement items for all constructs were adapted from prior research and designed to investigate the effects of device type and viewport dynamics. Measures of presence were adapted from the work of [19, 52]. Measures of motion sickness were adapted from [4]. Measures of audience engagement included three dimensions (behavioral, cognitive, and emotional engagement) and were adapted from [38, 46, 58]. Unless otherwise stated, all items were measured on a 7-point Likert scale (from 1 = strongly disagree to 7 = strongly agree). All measurement items are listed in the Appendix.

4 Results

4.1 Quantitative Analysis and Results

We used SPSS V.22 for our data analysis. To explore the differences among conditions, a one-way analysis of variance (ANOVA) was conducted. Table 3 details the ANOVA results for the independent variables (i.e., *Display Device*, *Viewport Dynamics*) with respect to presence, motion sickness, and audience engagement.

Table 3. ANOVA results

	Condition	M (SD)	F[1, 58]	p	η_p^2
<i>Presence</i>					
Display device	HMD	5.22 (.85)	5.45	.023*	.086
	Mobile	4.63 (1.09)			
Viewport dynamics	MVP	5.31 (.83)	9.88	.003*	.146
	SVP	4.54 (1.05)			
<i>Motion sickness</i>					
Display device	HMD	3.39 (.94)	11.63	.001*	.167
	Mobile	2.51 (1.07)			
Viewport dynamics	MVP	2.67 (.94)	4.05	.043*	.069
	SVP	3.23 (1.17)			
<i>Audience engagement</i>					
Display device	HMD	5.54 (.66)	5.56	.022*	.087
	Mobile	5.08 (.84)			
Viewport dynamics	MVP	5.58 (.61)	7.87	.007*	.120
	SVP	5.04 (.86)			

Note: * indicates $p < 0.05$

Main Effects

Presence. A simple main effect analysis showed that presence is significantly related to audience engagement ($p = .002$). An ANOVA of the independent variable *Display Device* regarding presence showed a significant difference between conditions ($F[1, 58] = 5.45, p = 0.023$), indicating that presence was higher for the HMD condition ($M = 5.22, SD = 0.848$) than the mobile device condition ($M = 4.63, SD = 1.09$). An ANOVA of the independent variable *Viewport Dynamics* regarding presence also showed a significant difference between conditions ($F = 9.88, p = 0.003$), including a greater presence for the MVP condition ($M = 5.31, SD = 0.828$) than the SVP condition ($M = 4.54, SD = 1.05$).

Motion Sickness. The analysis found a significant main effect of *Display Device* ($F = 11.63, p = 0.001$) for motion sickness. Specifically, participants reported a greater level of motion sickness in the HMD condition ($M = 3.39, SD = 0.937$) than in the mobile device condition ($M = 2.51, SD = 1.07$). We also found a significant difference in motion sickness between the MVP condition ($M = 2.67, SD = 0.936$) and the SVP condition ($M = 3.23, SD = 1.17$), ($F = 4.05, p = 0.043$). Participants who had viewed the videos with an MVP reported less motion sickness than those who had viewed the videos with an SVP.

Audience Engagement. We found a significant difference in audience engagement between the two types of display devices ($F = 5.56, p = 0.022$). The participants who used an HMD ($M = 5.54, SD = 0.659$) scored significantly higher on audience engagement than those who used a mobile device ($M = 5.08, SD = 0.844$). We also found that the factor *Viewport Dynamics* had a significant main effect on audience

engagement ($F = 7.87, p = 0.007$), showing that participants in the MVP condition ($M = 5.58, SD = 0.605$) reported higher audience engagement than those in the SVP condition ($M = 5.04, SD = 0.861$).

Indirect Effects

We conducted mediation analyses using Hayes's Model 4 [16] to determine the indirect effects of presence and motion sickness on audience engagement by applying 5,000 bootstrap samples and a 95% confidence interval (CI) [16]. The output showed two specific indirect effects.

Display Device. We found that presence was a significant mediator for audience engagement ($ab = 0.275, 95\% \text{ CI of the difference} = [0.038, 0.575]$). The results support H3b: Participants using HMDs have a higher sense of presence and thus higher engagement. We found no significant effect of motion sickness on audience engagement ($b = -0.123, p = 0.220$); therefore, H2 was not supported.

Viewport Dynamics. The effect of viewport dynamics on audience engagement was significantly mediated by presence ($ab = 0.351, 95\% \text{ CI of the difference} = [0.136, 0.616]$). The results suggest that the participants who watched the MVP video had a higher sense of presence and thus higher engagement. Motion sickness was found to have no significant effect on audience engagement; therefore, the mediation hypothesis (H5b) was not supported.

4.2 Qualitative Analysis and Results

We conducted a post-hoc analysis using qualitative feedback that was collected from the respondents regarding their video preferences. Two researchers coded the answers to the open-ended question using an iterative process of generating, redefining, and probing emergent themes. Themes adopted in the qualitative analysis included presence, viewport, interactivity, cognitive engagement, and emotional engagement.

Advantages and Disadvantages of MVPs

Advantages. Of the 60 (51.7%) participants, 31 mentioned the advantages of the MVP video. The coding results identified three primary reasons for why they preferred videos with an MVP to videos with an SVP: *presence* (14 of 31, 45%), *dynamic views* (11 of 31, 35.5%), and *interactivity* (9 of 31, 29.0%). Fourteen participants stated that they liked the MVP video because it increased their sense of presence. For example, participant 22 said, “[I]n the first one [MVP], I felt more like I was there with the dolphins.” Eleven participants identified the motion differences between the two videos and said that they preferred the MVP video. Participant 13 mentioned, “The first one was more dynamic and had more movement . . . I enjoyed moving through the rocks in the dust.” The participants’ preference for the MVP video was also due to its support of a greater sense of interactivity. Participant 38 stated, “I feel I can touch the animals, and the fishes in the video can watch me.”

Disadvantages. Of the 60 (30.0%) participants, 18 noted disadvantages of the MVP video. The top three cited disadvantages included *motion sickness* (7 of 18, 38.9%),

camera distance (5 of 18, 27.8%), and *cuts* (3 of 18, 16.7%). Seven participants stated that the MVP video caused more motion sickness. For example, participant 48 noted, “I was a little disoriented during the second video [MVP].” In addition, five participants reported not liking the camera distance in the MVP video. Participant 39 noted, “It seems objects in the video are at a far distance from me.” Finally, three participants expected smoother scene changes and a lower frequency of cuts in the MVP video. For instance, participant 15 mentioned, “[T]he second one [MVP] just has some sudden changes of view.”

Advantages and Disadvantages of SVPs

Advantages. Of the 60 (36.7%) participants, 22 mentioned the advantages of the SVP video. The top three cited advantages included *camera distance* (7 of 22, 31.8%), *enjoyment* (7 of 22, 31.8%), and *presence* (5 of 22, 22.7%). Seven participants preferred the SVP video because of camera distance. For example, participant 39 said, “I feel like sharks in the second video [SVP] are close to me.” Seven participants reported experiencing more enjoyment when watching the SVP video. Participants felt curious and interested in the SVP video: “The environment looks better and more interesting,” said participant 45. Finally, five participants reflected that the SVP video enhanced their sense of presence. Participant 40 said, “I certainly have a sense of ‘being there’; for me, the second one [SVP] is stronger.”

Disadvantages. Of the 60 (36.7%) participants, 22 reported disadvantages of the SVP video. Specifically, they cited the three main disadvantages as *SVP capabilities* (8 of 22, 36.4%), *motion sickness* (6 of 22, 27.2%), and *cuts* (5 of 22, 22.7%). Eight participants said that they were tired of the SVP. For example, participant 53 noted, “For the second one [SVP], mostly it just focuses on the object (i.e., the subject (me) does not move very often).” Similarly, participant 48 said, “[M]y vision in the second video [SVP] was stopped in a fixed position, so I could only watch from the same position.” In addition, six participants mentioned feeling dizzy when watching the SVP video versus the MVP video. Finally, five participants expected smoother scene changes and a lower frequency of cuts in the SVP video. For instance, participant 27 said, “I did not like the jump cuts in the second one [SVP].”

5 Discussion

The objective of this study was to assess the effects of display type and viewport dynamics on audience engagement with 360-degree videos. The results of the study suggest that device type and viewport dynamics jointly influence audience engagement through the mediation of presence. In this section, we discuss our key findings, implications, and possible directions for future research.

5.1 Findings

We found that audience engagement is mainly determined by the audience’s degree of presence while viewing a 360-degree video. Contrary to our expectation, motion

sickness was found to have an insignificant influence on audience engagement. The result is inconsistent with previous studies that highlight negative user experiences caused by motion sickness related to immersive technology [2, 28, 44]. One possible explanation is that a high degree of presence may override the negative effect of motion sickness on audience engagement. Previous studies have examined audience engagement from different perspectives by focusing on either presence or motion sickness. When we consider the positive (presence) and negative (motion sickness) viewing experiences simultaneously, our results show that presence may outweigh the negative effect of motion sickness on audience engagement.

In addition, our results show that different display devices generate different degrees of presence. It was found that HMDs outperformed mobile devices in creating greater degrees of presence. It was also found that HMDs increased motion sickness more than mobile devices. Previous studies have focused on the positive and negative experiences caused by HMDs. For example, some researchers have highlighted the positive effects of HMDs on presence [57], while others have highlighted the negative effects of HMDs on motion sickness [2]. Our study shows that HMDs cause both positive (presence) and negative (motion sickness) viewing experiences simultaneously.

Finally, we found that the extent to which an individual's perceived presence while viewing a video is influenced not only by display types but also by viewport dynamics. It was found that MVPs generate greater degrees of presence than SVPs. We conjecture that MVPs may create an illusion of interaction with video content and a more real experience than SVPs. We also found that videos with an MVP elicit less motion sickness than videos with an SVP. We conjecture that MVPs enable viewers to follow the motion of the viewport in a natural manner, while SVPs allow viewers to explore the video themselves by turning around quickly.

To summarize, our findings indicate that HMDs are more likely to increase presence than mobile devices do, and an MVP has an additive influence on presence. However, HMDs cause greater levels of motion sickness than mobile devices do. MVPs elicit less motion sickness and therefore offset the effect of HMDs on motion sickness.

5.2 Implications

Implications for Research. The present study provides the following key contributions to the literature on immersive technology. Because few studies on 360-degree videos have been conducted, this study can help researchers understand the current state of immersive technology research in terms of theoretical and methodological approaches, research themes, and contexts. This study also advances our knowledge on 360-degree videos by providing a comprehensive understanding of audience engagement. Most previous studies on 360-degree videos have focused on the effects of display types, while few have empirically investigated the effects of viewport dynamics on presence and motion sickness. Furthermore, although some researchers have discussed the importance of viewport dynamics in enhancing audience engagement, studies on display types and viewport dynamics have been conducted with disparate focuses and therefore do not inform each other. To the best of our knowledge, this

study was the first to combine the two factors as technological features of 360-degree videos and examine their joint effects on presence and motion sickness, which determine audience engagement. In so doing, the current study extends the existing literature on 360-degree videos. In addition, by illuminating the pathways from device type and viewport dynamics to audience engagement through presence and motion sickness, this study serves as a conceptual framework for audience engagement with 360-degree videos. Researchers can build on the proposed framework to develop new models that explain the interplay between technological features of 360-degree videos, audience experiences, and engagement. Finally, this study analyzed audience preference using a qualitative approach. Our findings regarding both the positive and negative audience experiences caused by display types and viewport dynamics will help researchers better understand both why and how audiences are engaged with 360-degree videos. Currently, 360-degree videos are incorporating virtual reality technology to enhance audience engagement in many contexts, including news media, marketing, and education. Our findings suggest that researchers who investigate the effects of immersive technology (i.e., virtual reality) should consider the fit between display types and viewport dynamics. Ignoring either of the two factors may lead to a failure to engage audiences. This study suggests that the joint effects of display types and viewport dynamics should be considered as unique aspects of 360-degree videos, which will benefit researchers who want to develop, test, and verify their theories regarding 360-degree videos.

Implications for Practice. The present study provides practical implications for system developers and managers who seek new ways to promote audience engagement through 360-degree videos. Our work revealed the interplay between different display types, presence, motion sickness, and audience engagement. Based on the findings, the fit between display types and viewport dynamics is key to enhancing audience engagement. If a 360-degree video is created using MVP, audiences should be encouraged to use HMDs. In case that a 360-degree video is created using SVP, audiences should be encouraged to use mobile devices while they view the video. Our findings also suggest that system designers should consider how to maximize audiences' sense of presence and minimize motion sickness by proposing the best combination between display types and viewport dynamics that can optimize their experiences.

5.3 Limitations

This study has some limitations. Although our experimental study is relevant for providing empirical results regarding 360-degree videos and its role in increasing audience engagement, we have identified limitations to measuring actual user experience and engagement. Researchers would benefit from considering method triangulation, for example, by employing neurophysiological measures, such as Electroencephalography, to detect a user's brain activity, which can help researchers to assess a user's mental state (e.g., relaxed or focused mental state) and the quality of user experience. Second, the data was collected from students of one university, which might lead to potential biases. In the future, the observed relationships should be tested in other population groups.

Acknowledgement. This research was supported by grant from the Centre for Applied Computing and Interactive Media (ACIM) of City University of Hong Kong awarded to the third author.

Appendix

Measurement Items

The items were measured on a 7-point Likert scale (from 1 = strongly disagree to 7 = strongly agree) unless otherwise stated.

Variable	Items	Source(s)
Presence	1. When watching the video, I had a sense of “being there.” (1: Not at all 7: Very much) 2. There were times during the experience when the video world became more real or present for me compared to the “real world” (1: At no time 7: Almost all of the time) 3. The video world seems to me to be more like (1. something that I saw 7. somewhere that I visited) 4. I had a sense of being in the video environment 5. I felt I was visiting the places in the video environment 6. I had a sense of being together with the objects in the video	[19, 52]
Motion sickness	While I was watching the video: 1. I felt dizzy looking at the screen 2. My eyes felt diplopia (double vision) 3. I felt blurred vision 4. I had a headache	[4]
Audience engagement	<i>Cognitive engagement</i> While I was watching the video 1. I lost track of the world around me 2. I blocked out things around me when I was watching the video 3. I was really drawn into the video 4. I felt involved in this video <i>Emotional engagement</i> 1. Watching the video was enjoyable 2. Watching the video was pleasant 3. Watching the video was interesting 4. Watching the video was fun <i>Behavioral engagement</i> 1. I intend to continue watching the 360-degree video in the future 2. I am willing to watch the 360-degree video in the future	[38, 46, 58]

References

1. Afzal, S., Chen, J., Ramakrishnan, K.: Characterization of 360-degree videos. In: Proceedings of the Workshop on Virtual Reality and Augmented Reality Network, pp. 1–6. ACM (2017)
2. Van den Broeck, M., Kawsar, F., Schöning, J.: It's all around you: exploring 360° video viewing experiences on mobile devices. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 762–768. ACM (2017)
3. Chirico, A., Cipresso, P., Yaden, D.B., Biassoni, F., Riva, G., Gaggioli, A.: Effectiveness of immersive videos in inducing awe: an experimental study. *Sci. Rep.* **7**(1), 1218 (2017)
4. Cho, S.-H., Kang, H.-B.: An assessment of visual discomfort caused by motion-in-depth in stereoscopic 3D video. In: *BMVC*, pp. 1–10 (2012)
5. Cochrane, T., Cook, S., Aiello, S., Christie, D., Sinfield, D., Steagall, M., Aguayo, C.: A DBR framework for designing mobile virtual reality learning environments. *Australas. J. Educ. Technol.* **33**, 54–68 (2017). Accepted for Special Issue on Mobile Augmented and Virtual Reality
6. Coxon, M., Kelly, N., Page, S.: Individual differences in virtual reality: are spatial presence and spatial ability linked? *Virtual Real.* **20**(4), 203–212 (2016)
7. De la Peña, N., Weil, P., Llobera, J., Giannopoulos, E., Pomés, A., Spanlang, B., Friedman, D., Sanchez-Vives, M.V., Slater, M.: Immersive journalism: immersive virtual reality for the first-person experience of news. *Presence Teleoperators Virtual Environ.* **19**(4), 291–301 (2010)
8. Diemer, J., Alpers, G.W., Peperkorn, H.M., Shiban, Y., Mühlberger, A.: The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Front. Psychol.* **6**(26), 1–9 (2015)
9. Ebbesen, M., Ahsan, S.: Virtual reality in experience marketing: an empirical study of the effects of immersive VR. *Marketing and Brand Management, Norwegian School of Economics* (2017)
10. Fernandes, A.S., Feiner, S.K.: Combating VR sickness through subtle dynamic field-of-view modification. In: 2016 IEEE Symposium on 3D User Interfaces (3DUI), pp. 201–210. IEEE (2016)
11. Ferris, S.H.: Motion parallax and absolute distance. *J. Exp. Psychol.* **95**(2), 258 (1972)
12. Fonseca, D., Kraus, M.: A comparison of head-mounted and hand-held displays for 360° videos with focus on attitude and behavior change. In: Proceedings of the 20th International Academic Mindtrek Conference, pp. 287–296. ACM (2016)
13. Freeman, J., Avons, S.E., Pearson, D.E., IJsselsteijn, W.A.: Effects of sensory information and prior experience on direct subjective ratings of presence. *Presence Teleoperators Virtual Environ.* **8**(1), 1–13 (1999)
14. Gugenheimer, J., Wolf, D., Haas, G., Krebs, S., Rukzio, E.: Swivrchair: a motorized swivel chair to nudge users' orientation for 360 degree storytelling in virtual reality. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 1996–2000. ACM (2016)
15. Hale, K.S., Stanney, K.M.: *Handbook of Virtual Environments: Design, Implementation, and Applications*. CRC Press, Boca Raton (2014)
16. Hayes, A.F.: *PROCESS: a versatile computational tool for observed variable mediation, moderation, and conditional process modeling*. White paper (2012)
17. Hettinger, L.J., Riccio, G.E.: Visually induced motion sickness in virtual environments. *Presence Teleoperators Virtual Environ.* **1**(3), 306–310 (1992)

18. Howarth, P., Costello, P.: The occurrence of virtual simulation sickness symptoms when an HMD was used as a personal viewing system. *Displays* **18**(2), 107–116 (1997)
19. Huang, T.-L., Hsu Liu, F.: Formation of augmented-reality interactive technology's persuasive effects from the perspective of experiential value. *Internet Res.* **24**(1), 82–109 (2014)
20. Huber, T., Paschold, M., Hansen, C., Wunderling, T., Lang, H., Kneist, W.: New dimensions in surgical training: immersive virtual reality laparoscopic simulation exhilarates surgical staff. *Surg. Endosc.* **31**(11), 4472–4477 (2017)
21. IJsselsteijn, W., de Ridder, H., Freeman, J., Avons, S.E., Bouwhuis, D.: Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence. *Presence Teleoperators Virtual Environ.* **10**(3), 298–311 (2001)
22. Kasahara, S., Nagai, S., Rekimoto, J.: First person omnidirectional video: system design and implications for immersive experience. In: *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 33–42. ACM (2015)
23. Ke, F., Lee, S., Xu, X.: Teaching training in a mixed-reality integrated learning environment. *Comput. Hum. Behav.* **62**, 212–220 (2016)
24. Kelaiah, I., Kavakli, M., Cheng, K.: Associations between simulator sickness and visual complexity of a virtual scene. *Frontiers* **3**(2), 27–35 (2014)
25. Kennedy, R.S., Stanney, K.M., Dunlap, W.P.: Duration and exposure to virtual environments: sickness curves during and across sessions. *Presence Teleoperators Virtual Environ.* **9**(5), 463–472 (2000)
26. Lee, S.H., Sergueeva, K., Catangui, M., Kandaurova, M.: Assessing Google Cardboard virtual reality as a content delivery system in business classrooms. *J. Educ. Bus.* **92**(4), 153–160 (2017)
27. Lin, J.-W., Duh, H.B.-L., Parker, D.E., Abi-Rached, H., Furness, T.A.: Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment. In: *Proceedings of IEEE Virtual Reality*, pp. 164–171. IEEE (2002)
28. Lin, Y.-C., Chang, Y.-J., Hu, H.-N., Cheng, H.-T., Huang, C.-W., Sun, M.: Tell me where to look: investigating ways for assisting focus in 360° video. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2535–2545. ACM (2017)
29. Linder, Å.: Key factors for feeling present during a music experience in virtual reality using 360 video. School of Computer Science and Communication, KTH Royal Institute of Technology (2017)
30. MacQuarrie, A., Steed, A.: Cinematic virtual reality: evaluating the effect of display type on the viewing experience for panoramic video. In: *2017 IEEE Virtual Reality (VR)*, pp. 45–54. IEEE (2017)
31. Magnus, U.: Navigating using 360° panoramic video: design challenges and implications. School of Natural Sciences, Södertörn University (2017)
32. McGill, M., Ng, A., Brewster, S.: I am the passenger: how visual motion cues can influence sickness for in-car VR. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 5655–5668. ACM (2017)
33. McMahan, R.P., Bowman, D.A., Zielinski, D.J., Brady, R.B.: Evaluating display fidelity and interaction fidelity in a virtual reality game. *IEEE Trans. Visual Comput. Graph.* **18**(4), 626–633 (2012)
34. Melo, M., Sampaio, S., Barbosa, L., Vasconcelos-Raposo, J., Bessa, M.: The impact of different exposure times to 360° video experience on the sense of presence. In: *Computação Gráfica e Interação (EPCGI), 2016 23° Encontro Português de*, pp. 1–5. IEEE (2016)

35. Muhammad, A.S., Ahn, S.C., Hwang, J.-I.: Active panoramic VR video play using low latency step detection on smartphone. In: 2017 IEEE International Conference on Consumer Electronics (ICCE), pp. 196–199. IEEE (2017)
36. Munafo, J., Diedrick, M., Stoffregen, T.A.: The virtual reality head-mounted display Oculus Rift induces motion sickness and is sexist in its effects. *Exp. Brain Res.* **235**(3), 889–901 (2017)
37. Narciso, D., Bessa, M., Melo, M., Coelho, A., Vasconcelos-Raposo, J.: Immersive 360° video user experience: impact of different variables in the sense of presence and cybersickness. *Univ. Access Inf. Soc.* 1–11 (2017)
38. O'Brien, H.L., Toms, E.G.: The development and evaluation of a survey to measure user engagement. *J. Assoc. Inf. Sci. Technol.* **61**(1), 50–69 (2010)
39. O'Brien, H.L., Toms, E.G.: What is user engagement? A conceptual framework for defining user engagement with technology. *J. Assoc. Inf. Sci. Technol.* **59**(6), 938–955 (2008)
40. O'Brien, H.L., Toms, E.G.: Examining the generalizability of the User Engagement Scale (UES) in exploratory search. *Inf. Process. Manag.* **49**(5), 1092–1107 (2013)
41. Ozcinar, C., De Abreu, A., Smolic, A.: Viewport-aware adaptive 360 video streaming using tiles for virtual reality. In: IEEE International Conference on Image Processing (2017)
42. Palmisano, S.: Consistent stereoscopic information increases the perceived speed of vection in depth. *Perception* **31**(4), 463–480 (2002)
43. Passmore, P.J., Glancy, M., Philpot, A., Roscoe, A., Wood, A., Fields, B.: Effects of viewing condition on user experience of panoramic video. In: Proceedings of the 26th International Conference on Artificial Reality and Telexistence and the 21st Eurographics Symposium on Virtual Environments, pp. 9–16 (2016)
44. Philpot, A., Glancy, M., Passmore, P.J., Wood, A., Fields, B.: User experience of panoramic video in CAVE-like and head mounted display viewing conditions. In: Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video, pp. 65–75. ACM (2017)
45. Prothero, J., Hoffman, H.: Widening the field of view increases the sense of presence within immersive virtual environments. Technical report, Virtual Environments Human Interface Technology Laboratory, University of Washington (1995)
46. Qiu, L., Bensabat, I.: Evaluating anthropomorphic product recommendation agents: a social relationship perspective to designing information systems. *J. Manag. Inf. Syst.* **25**(4), 145–181 (2009)
47. Rupp, M.A., Kozachuk, J., Michaelis, J.R., Odette, K.L., Smither, J.A., McConnell, D.S.: The effects of immersiveness and future VR expectations on subjective-experiences during an educational 360° video. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 2108–2112. SAGE Publications Sage CA, Los Angeles (2016)
48. Sánchez Laws, A.L.: Can immersive journalism enhance empathy? *Digit. Journal.* 1–16 (2017)
49. Sheikh, A., Brown, A., Watson, Z., Evans, M.: Directing attention in 360-degree video. In: IBC 2016 Conference (2016)
50. Shin, D., Biocca, F.: Exploring immersive experience in journalism. *New Media Soc.* **19**(11) 1–24 (2017)
51. Singla, A., Fremerey, S., Robitza, W., Raake, A.: Measuring and comparing QoE and simulator sickness of omnidirectional videos in different head mounted displays. In: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6. IEEE (2017)
52. Slater, M., Usoh, M., Steed, A.: Depth of presence in virtual environments. *Presence Teleoperators Virtual Environ.* **3**(2), 130–144 (1994)

53. So, R.H., Lo, W., Ho, A.T.: Effects of navigation speed on motion sickness caused by an immersive virtual environment. *Hum. Factors* **43**(3), 452–461 (2001)
54. Sundar, S.S., Kang, J., Oprean, D.: Being there in the midst of the story: how immersive journalism affects our perceptions and cognitions. *Cyberpsychol. Behav. Soc. Netw.* **20**(11), 672–682 (2017)
55. Sylaiou, S., Mania, K., Karoulis, A., White, M.: Exploring the relationship between presence and enjoyment in a virtual museum. *Int. J. Hum. Comput. Stud.* **68**(5), 243–253 (2010)
56. Tam, W.J., Stelmach, L.B., Corriveau, P.J.: Psychovisual aspects of viewing stereoscopic video sequences. In: *Stereoscopic Displays and Virtual Reality Systems V*, pp. 226–236. International Society for Optics and Photonics (1998)
57. Tse, A., Jennett, C., Moore, J., Watson, Z., Rigby, J., Cox, A.L.: Was I there?: impact of platform and headphones on 360 video immersion. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2967–2974. ACM (2017)
58. Venkatesh, V., Thong, J.Y.L., Chan, F.K.Y., Hu, P.J.-H., Brown, S.A.: Extending the two-stage information systems continuance model: incorporating UTAUT predictors and the role of context. *Inf. Syst. J.* **21**(6), 527–555 (2011)
59. Visch, V.T., Tan, E.S., Molenaar, D.: The emotional and cognitive effect of immersion in film viewing. *Cogn. Emot.* **24**(8), 1439–1445 (2010)
60. Von Der Pütten, A.M., Klatt, J., Ten Broeke, S., McCall, R., Krämer, N.C., Wetzel, R., Blum, L., Oppermann, L., Klatt, J.: Subjective and behavioral presence measurement and interactivity in the collaborative augmented reality game TimeWarp. *Interact. Comput.* **24**(4), 317–325 (2012)
61. Witmer, B.G., Singer, M.J.: Measuring presence in virtual environments: a presence questionnaire. *Presence* **7**(3), 225–240 (1998)
62. Zhou, C., Li, Z., Liu, Y.: A measurement study of Oculus 360 degree video streaming. In: *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 27–37 (2017)



Enhancing Bicycle Safety Through Immersive Experiences Using Virtual Reality Technologies

Hiroki Tsuboi, Shuma Toyama, and Tatsuo Nakajima^(✉)

Department of Computer Science and Engineering,
Waseda University, Tokyo, Japan
{ht918, toyama, tatsuo}@dcl.cs.waseda.ac.jp

Abstract. A bicycle is a fundamental means of transportation for citizens in a modern urban city. However, many traffic accidents have caused by bicycle riding. In this study, our aim is to make people learn desirable riding manners of a bicycle, and to improve their skills towards safer riding for bicycle riders by using a bicycle riding training systems based on virtual reality technologies. The system takes into account the following four issues. The first issue is improving the effectiveness of bicycle riding training by improving the reality of the training system. The second issue is the improvement of the awareness of bicycle riding safety by experiencing traffic accidents in the virtual space. The third issue is to introduce a proper feedback system to a user of the training system. The fourth issue is to introduce gamification-based rewards in the training system. We have developed a prototype system to validate our approach, and present some results from user studies to investigate the feasibility of the prototype training system.

Keywords: Augmented cognition technologies · Virtual reality
Bicycle riding · Avoiding physical risks

1 Introduction

A bicycle is a fundamental means of transportation for citizens in a modern urban city. However, many traffic accidents have caused by bicycle riding, for example, in Tokyo over 10,000 bicycle traffic injuries occurred in 2016 [6].

Our aim in this study is to make people learn desirable riding manners of a bicycle, and to improve their skills towards safer riding for bicycle riders by using virtual reality technologies.

One of advantages of using virtual reality technologies in the bicycle training system is to improve the training effect by offering immersive experiences about dangerous situations in a virtual space [1, 2]. In terms of offering the experience of dangerous situations, another advantage is that a user can experience traffic accidents without actual physical risks. The approach allows a user to experience more realistic experience, but makes it possible to make him/her predict dangerous situations without increasing real physical risks. For improving the bicycle training effect, the road condition offered in a virtual space should be close to the actual road condition as much

as possible, and his/her actual riding skills to ride on a bicycle should be improved in the virtual space.

In this paper, we first describe previous work on training system using virtual reality technologies in Sect. 2. Section 3 describes the design of the proposed training system developed in this study that focuses on four issues that we take into account in the design, and Sect. 4 explains the implementation of the prototype system. In Sect. 5, we present some observations of the initial prototype system that are necessary for designing the current version of the prototype system, then, Sect. 6 describes how this prototype system actually affects a user's safety riding consciousness, and finally in Sect. 6, we will present the conclusion of the paper.

2 Related Work

The effectiveness of driving Simulators has been reported in many studies[4, 5, 10]. Past researches like Psocka [7] showed that virtual reality technologies for education and training are very effective. They said that there are many difficulties in learning in terms of existing visual representations and simulations. Virtual reality technologies offer some possibilities to relieve these difficulties.

De Winter et al. claimed that a driving simulator has some advantages and disadvantages [3]. They said that driving simulators are advantageous in terms of reproducibility, standardization, the ease of data collection, and the possibility of encountering dangerous driving conditions without real physical risks to a user's body. On the other hand, low fidelity simulators can produce unrealistic driving and incorrect research results.

In our research, we introduce virtual reality technologies into a bicycle riding simulator and focus on the reality of user experiences, in a simulator, thus we believe that the main disadvantage of using a bicycle riding simulator to train people will be solved.

Schwebel and McClure [9] conducted children street-crossing training in a validated, interactive, and immersive virtual street environment. In this research, children received computer-generated feedbacks concerning safety immediately after every street-crossing. They said that virtual reality enables desirable training without the risk of physical injuries, and it provides a fun and an appealing environment for training.

Also, Wang et al. [11] introduced virtual reality technologies to improve the safety of bicycle riders. In the system used in their study, a user wears a head mounted display, and answers questions about his/her bicycle riding manners at each decision point in routes. This study concluded that training systems using virtual reality technologies are useful for children who are not familiar with bicycles.

On the other hand, in the bicycle training system we have developed, we especially focus on the reality of training in the system. For example, a user cannot move freely in a virtual space as shown in [11], whereas in our system, the user can move freely in the virtual space like real bicycle riding.

3 System Design

In this study, we consider the following four issues to build a more advanced virtual reality bicycle riding training system for improving a user's awareness on the safety of bicycle riders and for reducing bicycle accidents.

3.1 Improving the Reality of the Training System

The first issue is improving the effectiveness of bicycle training by improving the reality of the training system. We especially focus on the following two points.

Improve the Reality of Bicycle Operation

We created a bicycle training system with physical pedals and a bicycle handle bar. The system simply recognizes the movement of the pedals and the bicycle handle bar by using sensors. The system monitors the accelerations of the pedals, and the position of the bicycle handle bar, then the data are retrieved by a virtual bicycle in a virtual space, and the system allows a user to feel the reality to ride an actual bicycle.

More specifically, the moving speed of a bicycle in the virtual space is determined from the acceleration of its pedal obtained by an acceleration sensor and the current moving speed. Even if the pedal's rotation speed is the same, a way of the speed changes depends on the speed of the bicycle.

Also, the traveling direction of a bicycle in the virtual space is determined from the direction of the steering wheel and the current traveling direction. The direction of the travel does not reflect the direction of the handle of the bicycle instantaneously but gradually changes towards the direction of the steering wheel. This is because with the actual bicycle there is a time lag between when the handle bar operates the front wheels and when the entire body turns.

Improving the Reality of Visual Contents

In our bicycle training system, we used a 3D map that was realistically reproduced using the real city that ZENRIN Co., Ltd. has released by free of charge¹. In addition, by aligning the scale of the 3D map with the size of the city that was modeled in the training system (for example, the model size and the moving speed with respect to the height of a user's viewpoint), it makes it possible to demonstrate high reality.

Furthermore, as shown in Fig. 1, in order to reproduce an actual traffic situation, our system introduces a traffic light system, so that pedestrians move randomly in the map and basically they follow signals, or make a car travel around the map. The approach makes more sophisticated and realistic training possible.

Improving the Reality of Auditory Contents

There is a limitation to feel immersion in a training system through training that relies only on the human vision. Therefore, our system improves the immersion to the training system by also reproducing the running sound of a bicycle, the environmental sounds, and the shock sound at the time of an accident.

¹ <http://www.zenrin.co.jp/product/service/3d/asset/>.



Fig. 1. Traffic light and pedestrian system

3.2 Experiencing Accidents in the Virtual Space

The second issue is the improvement of the awareness of bicycle riding safety by experiencing realistic physical traffic accidents in a virtual space. The important advantage of using virtual reality technologies for bicycle training is to make a user to experience with traffic accidents without real physical injuries. Therefore, in this research, we propose a training system that makes a user strongly recognize potential risks when experiencing traffic accidents in riding a bicycle. The current design considers the following two points

Creating the Situations that Cause Traffic Accidents

We prepared two kinds of situations that cause accidents in the system. One is a careless accident experience (an accident experience that can be avoided if a user's care is taken not to cause an accident), and the other is an unavoidable accident experience. The accidental accident experience occurs when a user fails to pay his/her attention to safety. Specifically, it is a mechanism that an accident occurs when colliding with a randomly arranged pedestrians, cars, and others explained in Sect. 3.1. The unavoidable accident experience occurs mainly when a user puts out speed too much. Specifically, when a user enters an intersection at a high speed, a bike or a car collides with the user at such a speed that it cannot be avoided. In addition, as shown in Fig. 2, we also let people recognize the danger when the bike and car are crossing in front of them at a high speed.

Recognition of a Traffic Accident

In an accident experience using virtual reality technologies, it is necessary to make a user recognize that an accident has occurred with only the limited feeling without giving a shock to his/her body. We first evaluated the prototype training system that causes to move or rotate a user's field of view when an accident occurs and the user blows off in the virtual space. From the demonstration of the initial prototype system as



Fig. 2. A dangerous bike generated in the system

shown in Sect. 5, we found that it was hard to recognize that an accident occurs only by the visual effect, especially when a traffic accident occurred in a training system. Therefore, we modified the training system that makes it easier for a user to recognize his/her situations by adding not only visual expressions but also auditory expressions (ex. the impact noise due to a traffic accident).

3.3 Introducing a Proper Feedback System

The third issue is to introduce a proper feedback system. Normally, a bicycle rider cannot recognize whether his/her riding is safe or dangerous unless he/she meets actual accidents. Therefore, in this study, a user presents the moving data that he/she rides a virtual bicycle after he/she finishes his/her ride or when he/she has an accident. The proposed training system makes users understand whether the current traffic manners and riding safety are good or not through appropriate feedbacks.

In this feedback system, a user confirms his/her riding with both the bicycle rider's viewpoint and the bird's-eye viewpoint after riding a bicycle like in Fig. 3. By using the bird's-eye viewpoint, the user can check information on the user's blind spot that he or she often neglects his/her attention while riding a bicycle. Also, if he or she neglects a signal or approaches a pedestrian or a car too close, no alert will be issued during the training, but the feedback system will alert all of them and will lead to a better attention to users for the next time.

3.4 Introducing Gamification in the Bicycle Training System

The fourth issue is to introduce game mechanics based on gamification in the bicycle riding training system. We propose a method to implicitly induce a user to ride safely by using a scoring system. It is also intended to make it possible to recognize desirable traffic manners without prior explanation of desirable manners orally or in writing. We take into account the following two points in the current design.



Fig. 3. Feedback system window

In this research, we use some factors such as whether bicycle riding speed is not too fast, the distance to surrounding cars is appropriate, and whether signals are correctly observed as the indicator of safe riding. In addition, we use whether to pass the correct position on the road as an index of desirable traffic manners. In order to increase a user's bicycle riding skills, we adopt the scoring system that the score rises when ensuring these rules, and the score decreases when not ensuring these rules, as shown in Fig. 4. In addition, we also introduce a technique to obtain virtual items in a virtual space, like a green cross as shown in Fig. 5, whose scores rise substantially to the points where the scoring system likes to guide a user for navigating him/her towards desirable manners.



Fig. 4. Scoring system



Fig. 5. An item that gains the score (Color figure online)

Potential Pitfall of Gamification

There is a potential pitfall to use gamification that a user cannot correctly distinguish whether the action induced by the gamification is the actual safe bicycle riding and the correct bicycle riding manners or whether a user just follows a rule defined by the gamification mechanisms in the training system for increasing his/her scores. Due to this pitfall, there is a danger that training by the system has its effect only when there is a scoring system, and it will not be effective for actual riding in the real world. To solve this problem, we use the feedback system introduced in Sect. 3.3. By using the feedback system, a user can accurately grasp what was dangerous to his/her bicycle riding, so the user can understand the actual rules without, and he or she wants to follow the rules without the scoring system.

4 Implementation

Our training system consists of the following five components: a head mounted display (HTC Vive²), an HTC Vive Controller (a device that can be tracked the position), a simple exercise bicycle as shown in the left photo of Fig. 6, an acceleration sensor, and training system software. A user wears a head mounted display and is trained bicycle riding in a virtual city by pedaling and operating the handle of a simple exercise bicycle as shown in the right photo of Fig. 6.

Figure 7 shows the relationship of our system components. In our system, a user rides on a simple exercise bicycle and wears a head mounted display as described above. The HTC Vive Controller and the acceleration sensor are attached to a simple

² <https://www.vive.com/>.



Fig. 6. A simple exercise bicycle in the training system

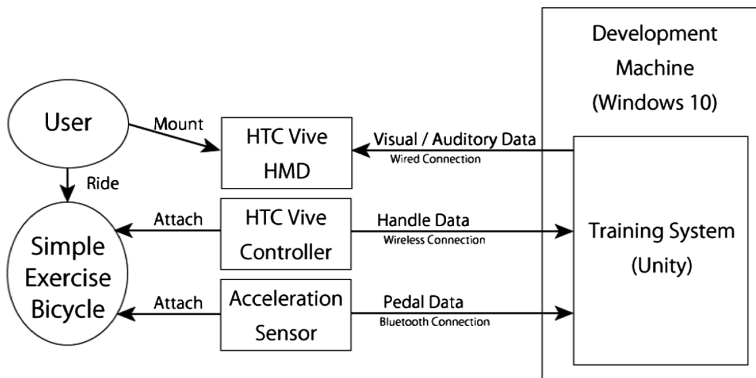


Fig. 7. System components

exercise bicycle and send the bicycle handle bars’ inclination and pedal acceleration data to the bicycle training system software. The training system displays a virtual city to a user based on the information.

The speed of the bicycle in a virtual space is determined from the rotational speed of a simple exercise bicycle pedal. The speed of the rotation of the pedals of a simple exercise bike is retrieved from the acceleration sensor contained in Nexus 6 (Android device), where the Nexus 6 is attached to the pedals.

The direction of the bicycle in the training system is determined from the inclination of the handle bars of simple exercise bicycle. The inclination is retrieved from the relative position information of the HTC Vive controller attached to the handle through its built-in sensors. The training system software has been developed on Unity (Version 2017.1.1), which is a cross-platform game engine primarily used to develop both three-dimensional and two-dimensional video games and simulations.

5 Demonstrating the Initial Prototype System in a University Event

In this section, we show some observations of the initial prototype system that have been developed before designing the current prototype system. The initial prototype has been developed for investigating potential pitfalls before developing the prototype systems described in the previous section, and make us extract the design issues described in Sect. 3. In an event held in our university where junior high school students and high school students participated as shown in Fig. 8, 170 users have experienced the initial prototype system. They could experience with virtual traffic accidents when traveling into an intersection of a road at the speed above a certain level.



Fig. 8. An event held in our university

Participants joined to the event gave us the following comments. A participant said *“I really feel like I’m riding on a bicycle”*. Some participants claimed *“Since a driving school like a car is not required for bicycle training, I feel that the training system is useful.”*, *“The reality offered by the training system is incomplete. For example, the shadow of a building in the virtual space is looked unnatural.”*, and *“It is hard to understand that I have met with an accident”*. Also, a participant told us *“I feel the drunkenness while using the initial prototype system.”*

6 Evaluating Learning Effects of the Prototype System

We conducted the experiment of the bicycle riding training for 10 participants, where 7 have the car driver’s license, and 3 do not have it. All 10 participants were in early twenties. In this experiment, we investigated how a participant’s riding behavior, skills

and attitudes were changed through using the training system before and after the training. Specifically, we examined whether the speed is not too fast, the distance to the surrounding cars and others is appropriate, the signals are followed, or whether a participant passes the correct spot on the road as an indicator. In addition to these points, we conducted a questionnaire survey and an examination on each participant’s awareness of safe bicycle riding and understanding of traffic laws.

6.1 Research Method

We conducted the experiment by the following steps.

- Step1:** Preliminary questionnaire
- Step2:** Training explanation
- Step3:** Actual training
- Step4:** Post questionnaire

In the bicycle riding training, each participant measured the lap to ride in the course as described before, and then watched the replay of his/her riding. After that, the participant rides the bicycle in the same course again one more time. If an accident occurs, the participant restarted from the place where the accident occurred and watches the replay of his/her riding later.

6.2 Result

Reality of the Training System

In the post questionnaire after the training, we investigated how participants felt the system’s reality with the 5-level Likert scale.

The results are shown in Fig. 9.

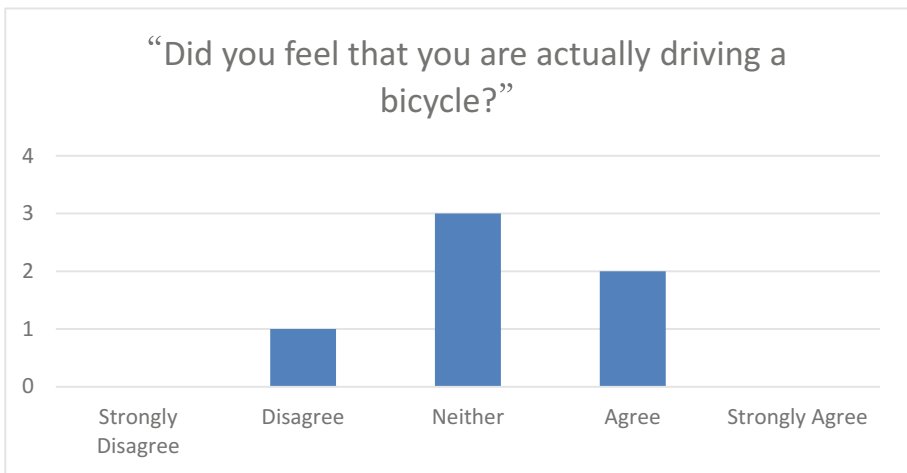


Fig. 9. Questionnaire on the reality of the training system

When asking what a reason that each participant felt the reality was low, he or she claimed that the problems are caused by the precision of the handle bar and the narrowness of the field of his/her view of the current prototype riding system. These problems are due to the current implementation's constraints not the concept's inherent limitation. Thus, we believe that these problems will be solved by the future system's improvements.

Bicycle Riding Behavior

In the experiment, we collected the information about the average speed, a number of signals ignored, a number of pedestrians approached, a number of collisions with pedestrians or cars, and a number of times a participant passed the places where he or she should not go through. We measured the differences between the data obtained in the first run and in the second run (after receiving some feedbacks to understand safe bicycle riding from the riding training system). During the replay of the participant's riding and the second run, the bicycle riding speed of his/her bicycle was indicated in the field of his/her view. At this time, the speed faster than the desirable speed was displayed for the purpose of preventing the participant's speeding out too much.

Unfortunately, no meaningful differences were found in the experiment. This may be caused by the short training time for avoiding VR sickness. Also most participants are aware of because they consider that they know that the purpose of the experiment is the training. For example, most participants ignored the signal only zero or one time in the first training, but some other participants decreased the number of signal ignoring in the second training.

Safety Awareness

We also conducted questionnaires on the participant's safety awareness before and after his/her bicycle training. In both questionnaires, we asked the following questions.

SQ1: What is necessary as a safety equipment (like a helmet)?

SQ2: What do you pay an attention to when riding a bicycle?

In the questionnaire before the bicycle riding training, we also asked the following questions.

EQ1: How do you recognize the danger of a bicycle?

EQ2: Do you forget to take care of the safety when riding a bicycle?

Then, after the bicycle riding training, we asked the following question.

EQ3: Do you think that your bicycle riding was safe so far?

Comparing before and after the experiment, SQ2 shows the difference. Many participants increased their attention during bicycle riding after training. Especially, a participant who answered "Not at all" or "Almost never forget" in EQ2 tends to pay more attentions.

Effect of Gamification

We confirmed to make a participant to recognize the correct traffic rules when using gamification. We chose a minor traffic rule (a bicycle must run on a bus lane in the road [8].) and designed the bicycle riding training system to give him/her a penalty if he/she does not keep the rule. We have checked the knowledge gain about the rules through questionnaires before and after the bicycle riding training.

In this experiment, we were able to find participants who were able to learn rules as intended. However, some participants misunderstood the rules. There is a difference between the understanding of the rules and their actual actions in such participants. So such participants' actual actions follow the rules.

7 Conclusion and Future Work

In this paper, we presented a bicycling riding training systems using virtual reality technologies. The system, in particular, took into account the following four issues. The first issue is improving the effectiveness of bicycle riding training by improving the reality of a training system. The second issue is the improvement of the awareness of bicycle riding safety by experiencing virtual accidents in a virtual space. The third issue is to introduce a proper feedback system. The fourth issue is to introduce gamification in the training system. We presented some results showing the effectiveness of the proposed approach from the user studies, and investigated its feasibility and some potential pitfalls of the current prototype system.

7.1 Future Work

Improving VR Device's Constraints

In this study, we developed a training system using general purpose commercial devices like an HMD and VR controllers. However, due to the limitations of the VR devices, some users have experienced VR sickness during their training, and felt the inconvenience in narrowing their views. We consider that these problems will be solved by using more sophisticated VR devices in the future.

Using Augmented Reality

In the future, some training functions developed in this study may be useful in actual bicycle riding in the real world. In particular, showing feedback information on the current riding is effective to understand the current riding manners. For example, displaying the speed faster than actual and showing warnings when pedestrians approach by using augmented reality technologies are useful for increasing safety of most bicycle riders when riding in the real world. Also, a scoring system showing scores by using augmented reality technologies in the real world is useful for bicycle riders for encouraging their safety manners through gamification.

Children's Training

Although we believe that this training system is the most effective for children who are weakly conscious about the safety and did not use a bicycle, it is pointed out that making children wear HMDs may have an adverse effect due to such as strabismus. So, we did not conduct an experiment with children in this study. However, if an HMD that can be used safely for children is developed, the training system proposed in this study will examine the effectiveness for the children.

References

1. Allen, R.W., Park, G.D., Cook, M.L., Fiorentino, D.: The effect of driving simulator fidelity on training effectiveness. In: DSC 2007 North America (2007)
2. Clancy, T.A., Rucklidge, J.J., Owen, D.: Road-crossing safety in virtual reality: a comparison of adolescents with and without ADHD. *J. Clin. Child Adolesc. Psychol.* **35**(2), 203–215 (2006)
3. De Winter, J., Van Leuween, P., Happee, P.: Advantages and disadvantages of driving simulators: a discussion. In: Proceedings of Measuring Behavior, pp. 47–50 (2012)
4. Fisher, D.L., Laurie, N.E., Glaser, R., Connerney, K., Pollatsek, A., Duffy, S.A., Brock, J.: Use of a fixed-base driving simulator to evaluate the effects of experience and PC-based risk awareness training on drivers' decisions. *Hum. Factors* **44**(2), 287–302 (2002)
5. Godley, S.T., Triggs, T.J., Fildes, B.N.: Driving simulator validation for speed research. *Accid. Anal. Prev.* **34**(5), 589–600 (2002)
6. National Police Agency Traffic Bureau: Traffic accident occurrence in 2016 (2017). <http://www.e-stat.go.jp/SG1/estat/Pdfdl.dl?sinfid=000031559551>. Accessed 12 Jan 2018
7. Pstoka, J.: Immersive training systems: virtual reality and education and training. *Instr. Sci.* **23**(5), 405–431 (1995)
8. Road Traffic Safety Management Office Environment and Safety Division, Road Bureau Ministry of Land, Infrastructure, Transport, and Tourism. Creating Safe and Secure Road Spaces for Cyclists. http://www.mlit.go.jp/road/road_e/pdf/Bicycle.pdf. Accessed 12 Jan 2018
9. Schewebel, D.C., McClure, L.A.: Using virtual reality to train children in safe street-crossing skills. *Injury Prevent.* **16**(1) (2010)
10. Underwood, G., Crundall, D., Chapman, P.: Driving simulator validation with hazard perception. *Transp. Res. Part F: Traffic Psychol. Behav.* **14**(6), 435–446 (2011)
11. Wang, W., Pratap Singh, K., (Mandy) Chu, Y.T.: Educating bicycle safety and fostering empathy for cyclists with an affordable and game-based VR App. In: MobileHCI 2016 Adjunct (2016)

Author Index

- Acosta, Eric II-293
Adler, Amy B. II-339
Ahmed, Alexis-Walid I-267, I-287, II-305
All, Anissa I-101
Almgren, Hannes I-101
al-Qallawi, Sherif II-205
Anzolin, Alessandra I-101
Auernheimer, Brent II-133
- Baltzer, Marcel C. A. I-9
Bang, Jounghae I-383
Barkan, Amanda I-59
Beckelheimer, Phillip I-267
Bernobić, Nikki I-287
Biddle, Elizabeth I-46
Biocca, Frank I-120
Blaha, Leslie M. I-245, II-43
Bombeke, Klaas I-101
Bovard, Pooja P. II-3
Bowles, Ben I-267
Boyce, Michael W. I-46, II-171, II-192
Bradascio, Joseph II-293
Brawner, Keith I-24
Brown, Payton I-267, II-255
Bruder, Gerd II-227
Brumback, Hubert K. II-15
Bryan, Derek S. II-363
Bryant, Andrew D. II-143
Burford, Clayton W. I-341
Burrell, Asher I-267
- Cardona-Escobar, A. F. I-158
Carlin, Alan I-24
Casebeer, William D. I-59, II-32
Choi, Y. Sammy I-395
Chun, Jaemin II-3
Cook, Kris II-43
Cope, Jamie II-293
Costello, Karen II-339
Cottam, Joseph II-43
Crawford, Chris I-212
Crosby, Martha E. I-316, II-117, II-133
Cummings, Mary L. II-154
Cunha, Meredith G. II-3
- Davis, Konrad L. II-326
Davis, Robert C. II-171
De Marez, Lieven I-101
DeFalco, Jeanine A. II-171, II-183
Dehais, Frédéric I-89
DeLellis, Stephen M. I-395
Deschamps, Anthony II-227
Dey, Anind K. II-3
Díaz, Gloria M. I-158
Djamasbi, Soussan II-105
Durnez, Wouter I-101
- Elfar, Mahmoud II-154
Elkin-Frankston, Seth II-58
Elliott, Linda R. II-67
- Fallon, Corey K. I-245
Feuerman, Jacob G. I-329
Fidopiastis, Cali II-214
Flemisch, Frank I-9
Forsyth, Carol II-143
Fortin-Côté, Alexis I-34
Frank, Gianella I-267
Fuchs, Sven I-3
- Gagnon, Jean-François I-34
Gallant, Scott I-341
Galvan-Garza, Raquel I-59
Garver, Sara K. II-3
Gilbert, Gary R. II-326
Girbacia, Florin I-170
Goldberg, Benjamin II-171, II-192
Gračanin, Denis I-355
Gray, Jeff I-212
Griffith, Richard L. II-205
Grigsby, Scott S. I-255
Gu, Wenying I-425
Guido-Sanz, Francisco II-227
Guo, Enruo II-143
- Hackett, Matthew II-240
Hadgis, Antoinette II-255
Hancock, Katy II-255

- Hancock, Monte I-267, II-305
 Hancock, Olivia I-267
 Hareide, Odd Sveinung II-273
 Harris, Tyler I-395
 Harrivel, Angela I-89
 Hedberg, Mathias I-369
 Helkala, Kirsi I-369
 Henry, Valerie II-293
 Hidalgo, Maartje I-341
 Hildre, Hans Petter II-350
 Ho, Nhut T. I-148
 Hoffmann, Lauren C. I-148
 Hollander, Markus I-287, II-255, II-305
 Hollingshead, Kristy I-180
 Hum, R. Stanley II-183
 Hurry, Mark I-406
- Ishihara, Manabu II-78
- Jaramillo-Garzón, J. A. I-158
 Jeon, Joonhyun I-120
 Johnston, Joan II-339
 Jøsok, Øyvind I-369
- Kanayama, Taiki II-78
 Kannan, Priya II-143
 Kennedy, Kellie I-89
 Kim, Gyoung I-120
 Kim, SeungJun II-3
 Knox, Benjamin J. I-369
 Kober, Erik K. II-171
 Kopf, Maëlle I-34
 Kow, Yong Ming I-406
 Kraft, Amanda E. I-59, II-32
 Kral, Daniel II-326
 Kramer, Diane I-24
- Lafond, Daniel I-34
 Lameier, Elizabeth I-46
 Larsen-Calcano, Tia II-94
 Lassen, Christian I-9
 Last, Mary Carolyn I-89
 Lee Van Abel, Anna I-148
 Lee, Jumin I-383
 Li, Charles I-267
 Lin, Jinchao I-131
 Liu, Alan II-293
 Liu, Wen II-105
 Lo, Chloe Chun-wing I-287
- Long, Rodolfo II-143
 Loos, Lila A. II-117
 López, Daniel I-9
 Lugo, Ricardo G. I-369
 Lützhöft, Margareta II-350
 Lyons, Joseph B. I-148
- Machidon, Octavian I-170
 Magee, J. Harvey I-395
 Mann-Salinas, Elizabeth II-326
 Marinazzo, Daniele I-101
 Marshall, Shana I-267
 Matthews, Gerald I-46, I-131
 Maymí, Fernando J. I-299
 McCracken, Kelsey II-240
 McDonald, Neil I-180
 McDonnell, Joseph I-329, I-341
 Meek, Wesley II-293
 Mercado, Gale I-267
 Milham, Laura II-339
 Miller, Geoffrey T. I-395
 Minas, Randall K. I-316, II-133
 Mogire, Nancy I-316, II-133
 Moon, Nicholas II-205
 Mortimer, Bruce J. P. II-67
- Nakajima, Tatsuo I-444
 Nelson, Kenneth I-395
 Neto, Nelson I-201
 Neumann, Shai II-305
 Nicholson, Denise II-227
 Niehaus, James II-58
 Nucci, Chris I-24
 Núñez Castellar, Elena Patricia I-101
 Nuon, Nick I-287
- Ochoa, Omar II-94
 Ogawa, Michael-Brian II-133
 Oster, Evan I-24
 Ostnes, Runar II-273, II-350
- Pajic, Miroslav II-154
 Pamplin, Jeremy C. II-326
 Pascarelle, Sebastian II-214
 Patton, Debbie II-339
 Pava, Matthew I-59
 Perez, Alison M. I-59
 Pérez-Zapata, A. F. I-158
 Perg, Lesley I-267

- Peters, Stephanie II-143
 Pettitt, Rodger A. II-67
 Poleski, Jason II-32
 Pope, Alan I-89
 Poquette, Melissa I-180
 Porras, Rainier A. I-329
 Postelnicu, Cristian-Cezar I-170
 Prophet, Jane I-406

 Quraishi, Nisha II-205

 Raj, Anil I-180
 Reichel, Howard II-214
 Reinerman-Jones, Lauren I-46, I-131, I-329,
 I-341, II-240
 Relling, Tore II-350
 Reyes, Fernando II-293
 Riddle, Dawn II-339
 Riecken, Mark E. I-341
 Roberts, Brooke I-180
 Roy, Raphaëlle N. I-89
 Russell, Bartlett II-32

 Sadler, Garrett G. I-148
 Sales Barros, Ellton I-201
 Salinas, Jose II-326
 Schlachta-Fairchild, Loretta II-326
 Schwartz, Jana L. II-3
 Shojaeizadeh, Mina II-105
 Shrider, Michael I-287, II-255
 Simonson, Richard II-94
 Sinatra, Anne M. I-69
 Skinner, Anna II-214
 Sood, Suraj II-305
 Sottolare, Robert I-78
 Soussou, Walid I-180
 Sprehn, Kelly A. II-3
 Start, Amanda R. II-339
 Steelman, Lisa A. II-205
 Stegman, Pierce I-212
 Stephens, Chad I-89
 Stiers, Frankie I-267
 Suh, Ayoung I-425

 Suh, Hyunju I-383
 Sütterlin, Stefan I-369

 Tanaka, Alyssa II-227
 Taylor, Glenn II-227
 Teo, Grace I-329, I-341, II-240
 Thomson, Robert I-299
 Townsend, Lisa II-339
 Toyama, Shuma I-444
 Trainor, Hayden J. I-329
 Trapp, Andrew C. II-105
 Tremblay, Sébastien I-34
 Tsuboi, Hiroki I-444

 Van Dongen, Aranka I-101
 Van Looy, Jan I-101
 Voinea, Gheorghe-Daniel I-170

 Wade, Rodney I-267
 Wagner, Christian I-425
 Walwanis, Melissa M. II-363
 Wan, Freda I-287, II-305
 Wang, Guan I-425
 Wang, Ziyao II-154
 Welch, Gregory II-227
 Whiting, Mark II-43
 Wilhelm, Michael II-183
 Wilkins, Mark I-148
 Williamson, Samuel I-267
 Willis, Sasha II-240
 Wismer, Andrew II-240
 Wohleber, Ryan I-131
 Wollocko, Arthur II-58
 Wooldridge, Robert E. II-67

 Yeaw, Ronald II-326

 Zapata-Rivera, Diego II-143
 Zhang, Rongrong I-222
 Zhao, Xiaojie I-222, I-231
 Zhu, Haipei II-154
 Ziegler, Matthias D. I-59, II-32
 Zuo, Shigang I-231