

Don Harris (Ed.)

LNAI 10276

Engineering Psychology and Cognitive Ergonomics

Cognition and Design

14th International Conference, EPCE 2017

Held as Part of HCI International 2017

Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part II

2
Part II



 Springer

Lecture Notes in Artificial Intelligence

10276

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>

Don Harris (Ed.)

Engineering Psychology and Cognitive Ergonomics

Cognition and Design

14th International Conference, EPCE 2017
Held as Part of HCI International 2017
Vancouver, BC, Canada, July 9–14, 2017
Proceedings, Part II

Editor
Don Harris
Faculty of Engineering and Computing
Coventry University
Coventry
UK

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-58474-4 ISBN 978-3-319-58475-1 (eBook)
DOI 10.1007/978-3-319-58475-1

Library of Congress Control Number: 2017939723

LNCS Sublibrary: SL7 – Artificial Intelligence

© Springer International Publishing AG 2017

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 Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

The 19th International Conference on Human–Computer Interaction, HCI International 2017, was held in Vancouver, Canada, during July 9–14, 2017. The event incorporated the 15 conferences/thematic areas listed on the following page.

A total of 4,340 individuals from academia, research institutes, industry, and governmental agencies from 70 countries submitted contributions, and 1,228 papers have been included in the proceedings. These papers address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The papers 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 on 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 2017 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.

April 2017

Constantine Stephanidis

HCI International 2017 Thematic Areas and Affiliated Conferences

Thematic areas:

- Human–Computer Interaction (HCI 2017)
- Human Interface and the Management of Information (HIMI 2017)

Affiliated conferences:

- 17th International Conference on Engineering Psychology and Cognitive Ergonomics (EPCE 2017)
- 11th International Conference on Universal Access in Human–Computer Interaction (UAHCI 2017)
- 9th International Conference on Virtual, Augmented and Mixed Reality (VAMR 2017)
- 9th International Conference on Cross-Cultural Design (CCD 2017)
- 9th International Conference on Social Computing and Social Media (SCSM 2017)
- 11th International Conference on Augmented Cognition (AC 2017)
- 8th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management (DHM 2017)
- 6th International Conference on Design, User Experience and Usability (DUXU 2017)
- 5th International Conference on Distributed, Ambient and Pervasive Interactions (DAPI 2017)
- 5th International Conference on Human Aspects of Information Security, Privacy and Trust (HAS 2017)
- 4th International Conference on HCI in Business, Government and Organizations (HCIBGO 2017)
- 4th International Conference on Learning and Collaboration Technologies (LCT 2017)
- Third International Conference on Human Aspects of IT for the Aged Population (ITAP 2017)

Conference Proceedings Volumes Full List

1. LNCS 10271, Human–Computer Interaction: User Interface Design, Development and Multimodality (Part I), edited by Masaaki Kurosu
2. LNCS 10272 Human–Computer Interaction: Interaction Contexts (Part II), edited by Masaaki Kurosu
3. LNCS 10273, Human Interface and the Management of Information: Information, Knowledge and Interaction Design (Part I), edited by Sakae Yamamoto
4. LNCS 10274, Human Interface and the Management of Information: Supporting Learning, Decision-Making and Collaboration (Part II), edited by Sakae Yamamoto
5. LNAI 10275, Engineering Psychology and Cognitive Ergonomics: Performance, Emotion and Situation Awareness (Part I), edited by Don Harris
6. LNAI 10276, Engineering Psychology and Cognitive Ergonomics: Cognition and Design (Part II), edited by Don Harris
7. LNCS 10277, Universal Access in Human–Computer Interaction: Design and Development Approaches and Methods (Part I), edited by Margherita Antona and Constantine Stephanidis
8. LNCS 10278, Universal Access in Human–Computer Interaction: Designing Novel Interactions (Part II), edited by Margherita Antona and Constantine Stephanidis
9. LNCS 10279, Universal Access in Human–Computer Interaction: Human and Technological Environments (Part III), edited by Margherita Antona and Constantine Stephanidis
10. LNCS 10280, Virtual, Augmented and Mixed Reality, edited by Stephanie Lackey and Jessie Y.C. Chen
11. LNCS 10281, Cross-Cultural Design, edited by Pei-Luen Patrick Rau
12. LNCS 10282, Social Computing and Social Media: Human Behavior (Part I), edited by Gabriele Meiselwitz
13. LNCS 10283, Social Computing and Social Media: Applications and Analytics (Part II), edited by Gabriele Meiselwitz
14. LNAI 10284, Augmented Cognition: Neurocognition and Machine Learning (Part I), edited by Dylan D. Schmorrow and Cali M. Fidopiastis
15. LNAI 10285, Augmented Cognition: Enhancing Cognition and Behavior in Complex Human Environments (Part II), edited by Dylan D. Schmorrow and Cali M. Fidopiastis
16. LNCS 10286, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Ergonomics and Design (Part I), edited by Vincent G. Duffy
17. LNCS 10287, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Health and Safety (Part II), edited by Vincent G. Duffy
18. LNCS 10288, Design, User Experience, and Usability: Theory, Methodology and Management (Part I), edited by Aaron Marcus and Wentao Wang

19. LNCS 10289, Design, User Experience, and Usability: Designing Pleasurable Experiences (Part II), edited by Aaron Marcus and Wentao Wang
20. LNCS 10290, Design, User Experience, and Usability: Understanding Users and Contexts (Part III), edited by Aaron Marcus and Wentao Wang
21. LNCS 10291, Distributed, Ambient and Pervasive Interactions, edited by Norbert Streitz and Panos Markopoulos
22. LNCS 10292, Human Aspects of Information Security, Privacy and Trust, edited by Theo Tryfonas
23. LNCS 10293, HCI in Business, Government and Organizations: Interacting with Information Systems (Part I), edited by Fiona Fui-Hoon Nah and Chuan-Hoo Tan
24. LNCS 10294, HCI in Business, Government and Organizations: Supporting Business (Part II), edited by Fiona Fui-Hoon Nah and Chuan-Hoo Tan
25. LNCS 10295, Learning and Collaboration Technologies: Novel Learning Ecosystems (Part I), edited by Panayiotis Zaphiris and Andri Ioannou
26. LNCS 10296, Learning and Collaboration Technologies: Technology in Education (Part II), edited by Panayiotis Zaphiris and Andri Ioannou
27. LNCS 10297, Human Aspects of IT for the Aged Population: Aging, Design and User Experience (Part I), edited by Jia Zhou and Gavriel Salvendy
28. LNCS 10298, Human Aspects of IT for the Aged Population: Applications, Services and Contexts (Part II), edited by Jia Zhou and Gavriel Salvendy
29. CCIS 713, HCI International 2017 Posters Proceedings (Part I), edited by Constantine Stephanidis
30. CCIS 714, HCI International 2017 Posters Proceedings (Part II), edited by Constantine Stephanidis

Engineering Psychology and Cognitive Ergonomics

Program Board Chair(s): **Don Harris, UK**

- Henning Boje Andersen, Denmark
- Martin Baumann, Germany
- Nicklas Dahlstrom,
United Arab Emirates
- Shan Fu, P.R. China
- John Huddleston, UK
- Kyeong-ah Kate Jeong, USA
- Wen-Chin Li, UK
- Andreas Luedtke, Germany
- Randy Mumaw, USA
- Jan Noyes, UK
- Paul Salmon, Australia
- Axel Schulte, Germany
- Patrick Waterson, UK
- Alf Zimmer, 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-2017.php>



HCI International 2018

The 20th International Conference on Human–Computer Interaction, HCI International 2018, will be held jointly with the affiliated conferences in Las Vegas, NV, USA, at Caesars Palace, July 15–20, 2018. 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 is available on the conference website: <http://2018.hci.international/>.

General Chair

Prof. Constantine Stephanidis

University of Crete and ICS-FORTH

Heraklion, Crete, Greece

E-mail: general_chair@hcii2018.org

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



Contents – Part II

Cognition and Design

System Latency Guidelines Then and Now – Is Zero Latency Really Considered Necessary?	3
<i>Christiane Attig, Nadine Rauh, Thomas Franke, and Josef F. Krems</i>	
Evaluation of Interface Modality for Control of Multiple Unmanned Vehicles.	15
<i>Gloria L. Calhoun, Heath A. Ruff, Kyle J. Behymer, and Clayton D. Rothwell</i>	
Research on User Mental Model Acquisition Based on Multidimensional Data Collaborative Analysis in Product Service System Innovation Process	35
<i>Jinhua Dou and Jingyan Qin</i>	
Are 100 ms Fast Enough? Characterizing Latency Perception Thresholds in Mouse-Based Interaction	45
<i>Valentin Forch, Thomas Franke, Nadine Rauh, and Josef F. Krems</i>	
Design and Evaluation of an Assistive Window for Soft Keyboards of Tablet PCs that Reduces Visual Attention Shifts	57
<i>Bomyeong Kim, Kyungdoh Kim, Jinho Ahn, and Robert W. Proctor</i>	
Integrated Information Visualization and Usability of User Interfaces for Safety-Critical Contexts	71
<i>Sonja Th. Kwee-Meier, Marion Wiessmann, and Alexander Mertens</i>	
The Study of Presentation Characteristics of the Warning Information and Its Influence on User’s Cognitive Process Based on Eye Tracking	86
<i>Yun Lin, Chengqi Xue, Qi Guo, Jing Zhang, Ningyue Peng, and Yafeng Niu</i>	
Cognitive Task Analysis for Interface Designs to Assist Medical Engineers in Hemodialysis Machine Troubleshooting	101
<i>Yoshitaka Maeda, Satoshi Suzuki, and Akinori Komatsubara</i>	
Design of a Decision-Making Task for a Collaborative Brain-Computer Interface System Based on Emotiv EEG	115
<i>Anderson Schuh and Márcia de Borba Campos</i>	

Effects of Key Size, Gap and the Location of Key Characters on the Usability of Touchscreen Devices in Input Tasks	133
<i>Da Tao, Qiugu Chen, Juan Yuan, Shuang Liu, Xiaoyan Zhang, and Xingda Qu</i>	
Natural, Multi-modal Interfaces for Unmanned Systems	145
<i>Glenn Taylor</i>	
UI-Design and Evaluation for Human-Robot-Teaming in Infantry Platoons	159
<i>Martin Westhoven, Christian Lassen, Irmtrud Trautwein, Thomas Remmersmann, and Bernd Brüggemann</i>	
“Smooth” or “Intermittent”? The Necessity of Halt in the Dynamic Visualization Due to the Features of Working Memory	179
<i>Xiaozhou Zhou, Chengqi Xue, An Li, Yafeng Niu, and Jing Zhang</i>	
Cognition in Aviation and Space	
Study on the Astronaut Error Criteria of a Manually Controlled Rendezvous and Docking Operation	191
<i>Jiayi Cai, Weifen Huang, Jie Li, Wang Liu, Haipeng Jing, Dong Chen, Yanlei Wang, and Xiang Zhang</i>	
Multi-modal Interaction Between Pilots and Avionic Systems On-Board Large Commercial Aircraft.	200
<i>Jason Gauci, Matthew Xuereb, Alan Muscat, and David Zammit-mangion</i>	
A Study for Human-Machine Interface Design of Spacecraft Display & Control Device Based on Eye-Tracking Experiments	211
<i>Qi Guo, Chengqi Xue, Yun Lin, Yafeng Niu, and Mo Chen</i>	
The Future Flight Deck	222
<i>Don Harris</i>	
Automated Online Determination of Pilot Activity Under Uncertainty by Using Evidential Reasoning	231
<i>Fabian Honecker and Axel Schulte</i>	
Assessing Human-Computer Interaction of Operating Remotely Piloted Aircraft Systems (RPAS) in Attitude (ATTI) Mode	251
<i>Pete McCarthy and Guan Kiat Teo</i>	
Multi-UAV Based Helicopter Landing Zone Reconnaissance: Information Level Fusion and Decision Support.	266
<i>Marc Schmitt and Peter Stütz</i>	

Factors Influencing Cargo Pilots’ Fatigue.	284
<i>Rui-shan Sun, Zi-li Chen, Guang-xia Huang-fu, Guang-fu Ma, Di Wu, and Zhen Liu</i>	
A Landing Operation Performance Evaluation System Based on Flight Data	297
<i>Lei Wang, Yong Ren, Hui Sun, and Chuanting Dong</i>	
Dynamic Measurement of Pilot Situation Awareness	306
<i>Xu Wu, Chuanyan Feng, Xiaoru Wanyan, Yu Tian, and Shoupeng Huang</i>	
An Approach for Assessing the Usability of Cockpit Display System	317
<i>Hongjun Xue, Tao Li, and Xiaoyan Zhang</i>	
Cognition and Driving	
Partial-autonomous Frenzy: Driving a Level-2 Vehicle on the Open Road . . .	329
<i>Francesco Biondi, Rachel Goethe, Joel Cooper, and David Strayer</i>	
The Human Element in Autonomous Vehicles	339
<i>Jerone Dunbar and Juan E. Gilbert</i>	
How Do Hybrid Electric Vehicle Drivers Acquire Ecodriving Strategy Knowledge?.	363
<i>Thomas Franke, Matthias G. Arend, and Neville A. Stanton</i>	
Design and Evaluation of a Mixed-Initiative Planner for Multi-vehicle Missions	375
<i>Fabian Schmitt, Gunar Roth, and Axel Schulte</i>	
A Field Study of Multimodal Alerts for an Autonomous Threat Detection System	393
<i>Erin T. Solovey, Pallavi Powale, and M.L. Cummings</i>	
Clustering of in-Vehicle User Decision-Making Characteristics Based on Density Peak	413
<i>Qing Xue, Qian Zhang, Xuan Han, and Jia Hao</i>	
Driver’s Multi-Attribute Task Battery Performance and Attentional Switch Cost Are Correlated with Speeding Behavior in Simulated Driving	426
<i>Jie Zhang, Mengnuo Dai, and Feng Du</i>	
Author Index	437

Contents – Part I

Mental Workload and Performance

A Method to Estimate Operator’s Mental Workload in Multiple Information Presentation Environment of Agricultural Vehicles	3
<i>Xiaoping Jin, Bowen Zheng, Yeqing Pei, and Haoyang Li</i>	
The Evaluation of Pilot’s First Fixation and Response Time to Different Design of Alerting Messages	21
<i>Wen-Chin Li, Jiaqi Cao, Jr-Hung Lin, Graham Braithwaite, and Matthew Greaves</i>	
An Analysis of Pilot’s Workload Evaluation Based on Time Pressure and Effort	32
<i>Wenmeng Liu, Yanyu Lu, Dan Huang, and Shan Fu</i>	
The Effects of Task Complexity and Spatial Ability on Teleoperation Performance.	42
<i>Dan Pan, Yijing Zhang, and Zhizhong Li</i>	
Model-Driven Payload Sensor Operation Assistance for a Transport Helicopter Crew in Manned–Unmanned Teaming Missions: Assistance Realization, Modelling Experimental Evaluation of Mental Workload	51
<i>Christian Ruf and Peter Stütz</i>	
Modeling of Performance Biases Induced by the Variance of Information Presentation to the Operator	64
<i>Sen Tian, Dan Huang, Lin Wang, and Shan Fu</i>	
Can Fixation Frequency Be Used to Assess Pilots’ Mental Workload During Taxiing?	76
<i>Xiaoyan Zhang, Hongjun Xue, Xingda Qu, and Tao Li</i>	

Psychological and Emotional Issues in Interaction

MINIMA Project: Detecting and Mitigating the Negative Impact of Automation	87
<i>Bruno Berberian, Oliver Ohneiser, Francesca De Crescenzo, Fabio Babiloni, Gianluca Di Flumeri, and Andreas Hasselberg</i>	
Cognitive Considerations in Auditory User Interfaces: Neuroergonomic Evaluation of Synthetic Speech Comprehension	106
<i>Adrian Curtin and Hasan Ayaz</i>	

Dynamic Changes of ERPs in Gestaltzerfall Phenomena: Analysis Using Multi-data Selecting and Averaging Method	117
<i>Mariko Funada, Tadashi Funada, and Yoshihide Igarashi</i>	
Decision-Making for Adaptive Digital Escape Route Signage Competing with Environmental Cues: Cognitive Tunneling in High-Stress Evacuation Situations	128
<i>Sonja Th. Kwee-Meier, Wolfgang Kabuss, Alexander Mertens, and Christopher M. Schlick</i>	
Factors Research on EEG Signal Analysis of the Willingness of Error Reporting.	141
<i>Hongxia Li and Nan Zhou</i>	
Mentally Imagined Item Captures Attention During Visual Search.	155
<i>Haifeng Li and Xiaomei Li</i>	
Evaluation of the Usability and Playability of an Exergame for Executive Functions Stimulation and Its Development Process.	164
<i>João Batista Mossmann, Eliseo Berni Reategui, Débora Nice Ferrari Barbosa, Rochele Paz Fonseca, Caroline de Oliveira Cardoso, and Vitor Caetano Silveira Valadares</i>	
Understanding the Relations Between Self-concept and Causal Attributions Regarding Computer Use	180
<i>Adelka Niels and Monique Janneck</i>	
Greater Heart Rate Responses to Acute Stress is Correlated with Worse Performance of Visual Search in Special Police Cadets.	200
<i>Xiaofang Sun, Yi Yuan, Zhuxi Yao, Kan Zhang, and Jianhui Wu</i>	
On-time Measurement of Subjective Anxiety of a Passenger in an Autonomous Vehicle: Gradually Changing Sounds Decreases Anxiety of Passenger	209
<i>Akitoshi Tomita, Etsuko T. Harada, Satoshi Ando, Kozue Miyashiro, Maito Ohmori, and Hiroaki Yano</i>	
Investigating the Influence of Emotion in Air Traffic Controller Tasks: Pretest Evaluation	220
<i>Martina Truschzinski, Georg Valtin, and Nicholas H. Müller</i>	
Stressor Load and Stress Resilience: A New Perspective for Occupational Stress	232
<i>Lijing Wang, Yanlong Wang, Yingchun Chen, Dayong Dong, and Wenjun Dong</i>	

Situation Awareness and Control

An Integrated Approach of Human Oriented Interactions with Complexity . . .	247
<i>Cedric Bach, Viviane Perret, and Guillaume Calvet</i>	
Human-Swarm Interaction as Shared Control: Achieving Flexible Fault-Tolerant Systems	266
<i>Jacob W. Crandall, Nathan Anderson, Chace Ashcraft, John Grosh, Jonah Henderson, Joshua McClellan, Aadesh Neupane, and Michael A. Goodrich</i>	
The Evaluation of Remote Tower Visual Assistance System in Preparation of Two Design Concepts	285
<i>Maik Friedrich, Stefan Pichelmann, Anne Papenfuß, and Jörn Jakobi</i>	
The Investigation Human-Computer Interaction on Multiple Remote Tower Operations	301
<i>Peter Kearney, Wen-Chin Li, Graham Braithwaite, and Matthew Greaves</i>	
Integrated Design of System Display and Procedural Display in Advanced NPP Control Rooms	310
<i>Yiran Ma, Qin Gao, Fei Song, and Yufan Wang</i>	
Design and Evaluation of an Abstract Auxiliary Display for Operating Procedures in Advanced NPP Control Rooms.	319
<i>Yahui Ma, Xiang Jiang, Qin Gao, Haitao Lian, and Qiuyu Wang</i>	
Authority Pathway: Intelligent Adaptive Automation for a UAS Ground Control Station	329
<i>Derek McColl, Kevin Heffner, Simon Banbury, Mario Charron, Robert Arrabito, and Ming Hou</i>	
An Evaluation of New Console Technology – Large Display – in Process Control Display	343
<i>Benjamin Noah, Jingwen Li, and Ling Rothrock</i>	
Use of Graphic Imagery as a Mean of Communication Between Operators and Unmanned Systems in C3Fire Tasks	362
<i>Tal Oron-Gilad and Ilit Oppenheim</i>	
Controller Intervention Degree Evaluation of Intersection in Terminal Airspace	382
<i>Yannan Qi, Xinglong Wang, and Xingjian Zhang</i>	
Implementation of a Responsive Human Automation Interaction Concept for Task-Based-Guidance Systems.	394
<i>Georg Rudnick and Axel Schulte</i>	

Team Situation Awareness: A Review of Definitions
and Conceptual Models 406
Manrong She and Zhizhong Li

Author Index 417

Cognition and Design

System Latency Guidelines Then and Now – Is Zero Latency Really Considered Necessary?

Christiane Attig¹(✉), Nadine Rauh¹, Thomas Franke², and Josef F. Krems¹

¹ Department of Psychology, Cognitive and Engineering Psychology,
Chemnitz University of Technology, Chemnitz, Germany
{christiane.attig,nadine.rauh,
josef.krems}@psychologie.tu-chemnitz.de

² Institute for Multimedia and Interactive Systems,
Engineering Psychology and Cognitive Ergonomics,
Universität zu Lübeck, Lübeck, Germany
franke@imis.uni-luebeck.de

Abstract. Latency or system response time (i.e., the delay between user input and system response) is a fundamental factor affecting human-computer interaction (HCI). If latency exceeds a critical threshold, user performance and experience get impaired. Therefore, several design guidelines giving recommendations on maximum latencies for an optimal user experience have been developed within the last five centuries. Concentrating on the lower boundary latencies, these guidelines are critically reviewed and contrasted with recent empirical findings. Results of the review reveal that latencies below 100 ms were seldom considered in guidelines so far even though smaller latencies have been shown to be perceivable to the user and impact user performance negatively. Thus, empirical evidence suggests a need for updated guidelines for designing latency in HCI.

Keywords: System response time · Latency · User experience · Design guidelines · Human-computer interaction

1 Introduction

Even though many technological advances aiming at fulfilling the quest for zero latency have emerged in recent years (e.g., regarding hardware and software speed, communication bandwidth), system latency still remains an inevitable aspect of human-computer interaction (HCI). If latency or system response time (SRT; i.e., the time interval between user input and system response), also known as lag or delay, exceeds a certain threshold, users are able to perceive and become aware of latency (e.g., [18]). If it increases even further, user experience (e.g., [35]) and satisfaction (e.g., [12]) can be impaired. Finally, also users' performance can be negatively affected by latency (e.g., [5]), even by latencies below the perceptual threshold [22].

For enabling system engineers and interface designers to create systems with the best user experience possible, several design guidelines for various applications have been established in the last 45 years (for overviews see e.g. [3, 9]). All these guidelines try to

answer the core question: Where are the latency thresholds? However, different guidelines for different aspects of HCI have to be distinguished. While some guidelines deal with human perception (e.g., what is the upper level of latency that users will just not notice?), others deal with user experience (e.g., what is the minimum latency where users start to get annoyed?). In this review, classic (i.e., before 1999) and more recent (i.e., since 2000) latency guidelines for designing interactive systems are examined. In the light of technical advances striving for zero-latency systems, our central question is: Are latencies close to zero considered necessary in these guidelines? Therefore, we concentrate on the lower latency limits that are specified in the reviewed latency guidelines (see Table 1).

Table 1. Latency guidelines and their lower limit latency recommendations.

Guideline	Smallest latency threshold	Characterization
Miller [23]	100–200 ms	<ul style="list-style-type: none"> • Latency guidelines for 17 different types of HCI • 100–200 ms is the longest acceptable latency for control activations • Based on the author’s expert estimation
Shneiderman and Plaisant [31]	50–150 ms	<ul style="list-style-type: none"> • Latency guidelines for different task complexity levels • 50–150 ms is the longest acceptable latency for basic, repetitive tasks • Based on empirical data
Card et al. [7]	100 ms	<ul style="list-style-type: none"> • Latency guidelines representing human perceptual limits • 100 ms is the maximum latency for creating the illusion that a system runs instantaneously • Generalized from classic psychophysical experiments
Seow [29]	100–200 ms	<ul style="list-style-type: none"> • Latency guidelines for different user expectations • 100–200 ms is the longest acceptable latency for system responses that the user expects to be instantaneous • Foundations not clearly stated
Tolia et al. [32]	150 ms	<ul style="list-style-type: none"> • Latency guidelines for interactions with thin clients • Below 150 ms user performance will not be negatively influenced and the user will not notice the latency • Based on previous guidelines and empirical data
Kaaresoja et al. [19]	visual: 30–85 ms audio: 20–70 ms tactile: 5–50 ms	<ul style="list-style-type: none"> • Latency guidelines for different feedback modalities after touchscreen button presses • Perceived button quality will decrease with latencies above the thresholds • Based on empirical data
Kaaresoja [20]	visual-audio visual: 90 ms audio: 70 ms visual-tactile visual: 100 ms tactile: 55 ms tactile-audio tactile: 25 ms audio: 100 ms	<ul style="list-style-type: none"> • Latency guidelines for bimodal feedback after touchscreen button presses • Perceived button quality will decrease with latencies above the thresholds • Based on empirical data
Doherty and Sorenson [11]	300 ms	<ul style="list-style-type: none"> • Latency guidelines for different user expectations and attentional states • Below 300 ms the users will feel as if they are in direct control • Based on previous guidelines and empirical data

2 Classic Latency Guidelines

The first author to determine latency thresholds was Miller in 1968 [23]. His design recommendations for various types of HCI were based on “the best calculated guesses by the author” ([23], p. 271), that is, they were not based on systematic empirical investigations (see also [5]). These early guidelines, which were focused on user acceptance (i.e., acceptable latencies), were theoretically grounded on two pillars: (1) common expectancies in interpersonal communication (i.e., typical patterns of interpersonal communication) and (2) memory research. Regarding the first aspect, according to Miller, in a conversation between two people an answer is expected within a few seconds. If the response delay exceeds four seconds, the thread of communication breaks [23]. Miller applied this pattern to HCI, which he viewed as a conversational act similar to a dialogue between two people, and defined maximum SRTs for 17 different kinds of conversational acts between the user and the system. Regarding the second aspect, due to the limited capacity of short-term memory, human thought and problem solving processes are interrupted if the SRT exceeds a certain threshold. The longer a chunk has to be kept active in short-term memory, the more likely are the chances of errors or forgetting (e.g., the chances of forgetting an e-mail address rise with increasing delay in loading the e-mail software). According to Miller, the longest acceptable latency for the system response in the most basic interactions (control activations, i.e., feedback that signals physical activation, e.g., an audible mouse click) is 100–200 ms. SRTs below 100 ms are not mentioned by Miller. Being aware that his recommendations can only be a starting point, Miller urged the need for empirical validation of his guidelines. Nevertheless, they constituted a first valuable guidance for practitioners and were used as reference in research on SRT and its effects on user experience.

In his review on SRT and human performance, Shneiderman [30] summarized experimental research on SRT and underlined the importance of users’ expectations for the acceptance of latencies. Expectancies are influenced by three factors [30, 31]: (1) prior experience, (2) an individual’s tolerance for and adaptability to delays, and (3) task complexity. First, prior experience with a certain kind of task shapes a user’s expectations regarding the same or similar tasks in the future (e.g., if a user learns that the delay between a search query in Google and the display of results is 300 ms, s/he will expect future search processes to take the same amount of time). Second, several person variables (e.g., age, professional experience, mood) determine a user’s willingness to wait. Moreover, people can adapt to long SRTs (e.g., by fulfilling other tasks while waiting). Third, with increasing task complexity, users are willing to accept longer SRTs. An experiment investigating simple, repetitive control tasks [15], which Shneiderman [30] referred to, showed SRTs below 1 s (i.e., 160 ms, 720 ms) to be superior for user performance (in contrast to 1149 ms). Regarding more complex problem solving tasks, the picture is less clear: While users had a more favorable attitude towards a low-latency system (330 ms), they made fewer errors with a longer latency (1250 ms; [33]). Furthermore, the higher the complexity, the higher users’ adaptation to the latency [30]. In sum, for simple and repetitive tasks, users have a higher satisfaction and better performance if SRTs are short. In contrast, users can adapt to longer SRTs in complex tasks, but their satisfaction decreases with increasing SRT [30, 31]. Based on these empirical results,

Shneiderman and Plaisant [31] defined task-centered latency guidelines regarding user acceptance for tasks with different complexity levels. According to the authors, the most basic, repetitive tasks (e.g., single keystrokes and mouse clicks) require SRTs from 50-150 ms to keep the user satisfied. However, the theoretical basis for the lower boundary of 50 ms remains unclear. Moreover, it is not explicitly stated for which kind of tasks latencies below 100 ms are required, thus, it can only be assumed that users with high prior task experience and a low tolerance for delays prefer very small latencies in simple tasks (i.e., below 100 ms). Yet, as Dabrowski and Munson [9] point out, a definition of task complexity is missing in Shneiderman's classification, thus, it remains unclear what exactly makes a task complex.

Choosing a different approach, Card, Robertson, and Mackinlay [7] referred to psychophysical experiments investigating human perception thresholds (e.g., regarding apparent motion; [6]) and applied those results to HCI. According to the authors, for creating the illusion that a system runs instantaneously, a maximum SRT of 100 ms has to be applied, otherwise the user will notice the delay (e.g., distinct lights on a graphical user interface instead of a single light in motion; [6])¹. This 100 ms threshold of perceptual processing was later made popular by Nielsen ([24]; see also [29]). Together with the early work by Miller [23], the work of Card et al. [6, 7] made the 100 ms threshold a frequently cited design rule implying that longer SRTs are not acceptable to the user [27].

However, in the 100 ms rule of thumb empirical data regarding perceptual thresholds [6] and subjective estimates regarding user acceptance [23] are somehow entangled. In guidelines based on empirical data regarding user latency acceptance also latencies below 100 ms are mentioned, at least for the most basic computer tasks [31]. Nevertheless, as we will see in the next section, 100 ms remained the lower bottom SRT guideline even in modern design guidelines, implying that SRTs below this threshold should not affect users markedly.

3 Recent Latency Guidelines

In his book on time perception in HCI, Seow [29] emphasized the importance of user expectations for establishing latency guidelines. He stated, similar to Shneiderman [30], that latency acceptance is relative to users' expectations and the nature of the task (i.e., longer latencies are acceptable for tasks with higher complexity as these are expected to require more computing capacity, and therefore, more time). In contrast to Shneiderman [31], he did not derive guidelines for different levels of task complexity but for different user expectations (i.e., instead of task-centered, his guidelines are user-centered with a stronger focus on the interaction). According to Seow [29], users have certain expectations regarding the responsiveness of the system if a certain task is conducted. For instance, tasks that mimic events in the physical world with instantaneous responses (e.g., pressing a virtual button which mimics

¹ It has to be emphasized that Card et al. referred to classic experiments investigating apparent motions. In these, influences of different framerates – and not input latency – on human perception were investigated.

pressing a physical button) should also show instantaneous responses (e.g., an audible click). For this very basic kind of task, the user expects the system to respond instantaneous, which means that a maximum SRT of 100 ms is required for very simple feedback (e.g., audible click after a virtual button press), respectively 200 ms for slightly more complex feedback (e.g., visual drop down menu). The next category, labelled “immediate”, concerns situations in which the user expects the system to respond by performing an action initiated by the user (e.g., the display of a letter after a keystroke) and requires a maximum SRT of 500–1000 ms [29]. It remains unclear on what data these latency thresholds are grounded on as no empirical data are presented.

Different from these universal guidelines, some guidelines for single use cases have been developed. Tolia, Andersen, and Satyanarayanan [32] defined latency guidelines for thin clients (i.e., lightweight computers using remote access to a server to run applications). In this case, besides the latency within the application, the end-to-end communication from user to server and back produces additional latency. This is a particular challenge for system engineers, because users are nowadays used to systems without perceivable delay [32]. Based on prior empirical work and latency guidelines [23, 31], the authors concluded that user performance is not negatively influenced by SRTs below 150 ms. Therefore, in order to perceive the thin client’s system output as immediate, the SRT (here: end-to-end latency meaning the time it takes from user input to server and back until the display of system output) must not exceed 150 ms, otherwise, the delay will get noticeable (>150 ms) and, finally, the interaction becomes annoying (>1000 ms). Thus, this guideline contains recommendations both for latency perception and user experience.

In contrast, Kaaresoja, Brewster, and Lantz [19] made a clear distinction between perception and user experience by empirically investigating both variables independently and deriving latency guidelines for another specific use case: touchscreen button presses. By experimentally manipulating the latency between the first finger touch and system feedback as well as feedback modality (visual, audio, tactile), the authors calculated the point of subjective simultaneity (PSS) for each feedback modality and, in addition, assessed users’ perceived quality of the touchscreen button. Combinations of the three different feedback modalities and nine different latency conditions (ranging from 0 to 300 ms, in addition to the baseline system latency) were presented. Users had to state if the feedback appeared simultaneously with their touch and, in a later but similar phase, how s/he would rate the quality of the button (from 1 = low quality to 7 = high quality). It was reported that the PSS for visual feedback was 32 ms, for audio feedback 19 ms and for tactile feedback 5 ms. Thus, the participants were able to perceive very small latencies, especially for tactile feedback. Significant drops in the perceived quality scores were found at 100–150 ms for visual, and 70–100 ms for audio as well as for tactile feedback. Moreover, buttons with any feedback with a 300 ms latency were rated significantly lower than the buttons with any feedback with latencies ranging from 0 to 150 ms. According to the guidelines by Kaaresoja et al. [19], latencies for visual feedback should lie between 30–85 ms, for audio feedback between 20–70 ms and for tactile feedback between 5–50 ms. Hence, their guidelines were the first to explicitly incorporate latencies smaller than 50–100 ms, if only for a very specific use case.

Using a similar experimental approach, Kaaresoja [20] expanded his guidelines for bimodal feedback (i.e., visual-audio, visual-tactile and tactile-audio). It was found that for different feedback pairs different levels of symmetry between the two feedback modality latencies emerge, as follows. For the combination of visual and audio feedback, the visual feedback latency should not be greater than 90 ms while the audio feedback should not exceed 70 ms. For the combination of visual and tactile feedback, the visual feedback latency should not be greater than 100 ms while the tactile feedback should not exceed 55 ms. And lastly, for the combination of tactile and audio feedback, the tactile feedback latency should not be greater than 25 ms while the audio feedback should not exceed 100 ms.² The empirical results [19, 20] suggest a high sensitivity for delay of tactile feedback in tactile HCI. This finding is in line with the suggestion that interactions which mimic events in the physical world (e.g., tactile feedback after virtual button touch) require very small latencies to be perceived as instantaneous [28].

In their review, Doherty and Sorenson [11] updated and expanded the existing general latency guidelines [29, 30] with a special focus on the flow experience [8]. The authors argue that in the usage of today's frequently used interactive systems (e.g., smartphones, tablets) short interactions (e.g., menu navigation, scrolling) are predominant. As it has been pointed out before, small latencies will get noticed or even annoy the user especially in very short and basic interactions (e.g., [19, 23, 29, 31, 32]). One negative influence of perceived latency is that users' interaction with the system can be interrupted, thus, users' flow gets broken ([11]; see also [29]). Incorporating empirical results on user expectations, perceived task complexity and perceptual limits, Doherty and Sorenson's guidelines [11] represent the most elaborate latency guidelines for an optimal user experience so far. However, the authors raised the lower boundary latency threshold for instantaneous responses to 300 ms. This figure was incorporated because of Kaaresoja's [20] finding that the perceived quality of the touchscreen button was significantly lower with 300 ms latency in contrast to 0–150 ms. Thus, "[...] depending on the input modality (mouse, keyboard, touchscreen, air, gesture, speech, etc.), the perception of what a user would consider instantaneous will vary." ([11], p. 4390). While the lower limit of 300 ms gives the guideline a higher generalizability, it also decreases its accuracy for very short interactions.

It becomes apparent that latencies below 100 ms do not play a role in most design guidelines. The only general guideline that explicitly mentioned a latency threshold smaller than 100 ms was the one by Shneiderman and Plaisant [31], but it was not explicitly stated under which conditions (e.g., task demands, user status) a latency has to be as small as 50 ms to be acceptable. The only other guideline recommending maximum latencies below 100 ms is the one by Kaaresoja et al. [20], suggesting that in very basic interactions (i.e., control tasks; [9]) – the ones that Miller [23] called "control activations" and Seow [29] expected to be "instantaneous" – user experience gets significantly impaired by latencies below 100 ms. Still, following the majority of guidelines, zero-latency systems do not seem necessary for optimal user experience. But is this really the case?

² Note that the dependent variable was a PSS judgment. Thus, based on these results, users will notice the delay if one of the feedback modalities exceeds the latency thresholds. If and when the participants perceived an asynchronicity between the two feedbacks was not assessed.

4 Empirical Evidence for the Perception of Latencies Below 100 ms and their Impact on HCI

Within recent years, there has been a considerable growth of studies examining latency effects in HCI even below the 100 ms threshold, possibly also because of increasing technical potentialities (e.g., high-speed cameras). In several studies, system latency was experimentally varied and perceptual limits were tested by applying classic psychophysical methods (i.e., estimating the just noticeable difference regarding perceived latency between two identical tasks with different latencies). These studies, which are presented in the following, indicate that users are indeed able to perceive latencies well below 100 ms. In addition, other studies show that even such small latencies can have negative effects on user performance – even when the latencies are below the perceptual threshold. Moreover, influencing factors on the perception of latencies are investigated, implying that latency perception is dependent on user and task variables.

During a digital inking task using a stylus [2], users were able to perceive latencies between input (i.e., the touch of the stylus on the screen) and visual feedback (i.e., the appearance of the digital ink) down to 50 ms with slightly higher perception thresholds for tasks that require more attentional resources (i.e., cause a greater workload). In a direct dragging task on a touchscreen, users were even able to notice latencies down to 11 ms [10], 6 ms [27] and even down to 2 ms under specific circumstances [25]. And even in a direct tapping task on a touchscreen (i.e., button press) where relatively few data are available to make latencies salient, a perceptual threshold of 64 ms was found [18].

So far, these results all refer to zero-order tasks. Zero order is one type of control order, that is, the way that the system responds to a change of the position of the control [34]. In zero-order tasks, a change in the position of the control (e.g., the mouse on the mousepad) leads to a change in the position of the displayed system output (e.g., the cursor on the screen; [17, 34]). In contrast, first-order control tasks require velocity control [34]. Here, a change in the control position leads to a constant change of velocity (e.g., a button press on a DVD remote control to raise up the playback speed to 2x). Finally, second-order control tasks deal with a change of acceleration (i.e., changes in the rate of velocity) and require more cognitive resources than zero- and first-order control tasks. One example in the field of vehicle control is the relationship between steering wheel position and the vehicle's lateral position in the lane. Here, a constant change in the steering wheel position leads to an increasing rate of change in the lateral position [34]. In second-order tasks, when the input is set to zero, the output continues to change and is not instantly set to zero as it is the case in zero- and first-order tasks [17]. Such a more demanding, second-order task was applied in an own study [22]. Using a virtual balance task, it has been shown that performance was already impaired by an added latency of 49 ms (technical base latency: 10.8 ms). However, participants perceived only the added latency from 97 ms on. Hence, even though users were not able to perceive the latency, it had an effect on their performance.

The effect of latency on user performance was also examined more closely in recent years. For instance, Brady et al. [4] applied an indirect mouse movement task and found that an added latency of 33 ms significantly impaired user performance.

In a pointing task, latency began to affect performance at 16 ms [14]. In a 3D game environment, a latency of 41 ms impaired user performance in an aiming task [16].

5 Factors Affecting Latency Perception

The studies presented so far were concerned with identifying latency thresholds for perception and performance. Other studies examined effects of influencing factors on latency perception, suggesting that latency thresholds are not cast in stone, yet, are system-, task- and person-dependent. Hence, a key task from an engineering psychology perspective is to structure relevant variables affecting latency perception. In the following, empirical results as well as assumptions regarding (1) system characteristics, (2), task characteristics, and (3) person characteristics will be discussed.

First, concerning system characteristics, different *input modalities* can be distinguished. When comparing direct (e.g., via touchscreen) and indirect input (e.g., with conventional input devices such as a mouse), sensitivity to latencies is higher in direct interaction [10, 26]. This can likely be ascribed to a higher salience of the latency because the visual attention is located within the same place as the system input. Another factor is the *output modality*: The latency perception thresholds differ with respect to the modality of the feedback after a virtual button press. Users are extremely sensitive to a latency in tactile feedback (when compared to audio and visual feedback) when the input is also tactile [19, 20]. According to Seow [29], a tactile feedback after a virtual button press is very similar to the press of a real physical button, therefore the user expects an instantaneous response and might be more sensitive to interaction delays. Moreover, the *number of feedbacks* seems to play a role in latency perception. When two feedbacks are provided in contrast to just one, latency sensitivity is lower [18, 20, 28]. One explanation for this effect might be an additional information-processing step which is needed to integrate the two feedbacks [25], however, this remains speculative at the present time. In visual dragging tasks, the *size ratio between physical reference and visual feedback* affects latency perception. If the size of the physical reference (e.g., a stylus nib) and the visual feedback are more similar, latency perception is improved [25]. Possibly this can also be attributed to the higher similarity to an interaction in the physical world [29].

Second, regarding task characteristics, an important factor that has already been incorporated in guidelines is *task complexity*. By experimentally varying task complexity, two studies found that users perceive smaller latencies in simple tasks (i.e., dragging tasks) compared to slightly more complex, thus, demanding tasks (i.e., scribbling tasks; [2, 25]. Moreover, *interaction speed* affects latency perception in dragging tasks. The faster the user's hand motion in a dragging task, the better the latency perception [26]. This finding is attributed to the visual effect of a fast hand motion in a dragging task which creates the illusion that the displayed square is "attached to a rubber band to the user's finger" ([26], p. 453). This effect makes latency visible and salient to the user. The latency perception model [1], which describes the process of latency perception, postulates that the user utilizes a referent to make latency judgments. More specifically, a referent is a stimulus within the interaction (e.g., a stylus nib, the user's finger) that

the user compares to the system response to evaluate the latency magnitude [1]. One example is the user's hand in a dragging task as described before [26]. According to [1], the *presence of a referent* affects latency perception. If the hand is made invisible and can therefore not be used as a referent, latency sensitivity is diminished in a scribbling task [2]. Further, the *modality of the referent* is discussed as a factor influencing latency perception [1].

Finally, regarding person characteristics, domain specific *experience* seems to be an important factor for the perception of latencies. The experience with highly dynamic computer games (i.e., action games, racing games, first person shooter games) was found to correlate positively with latency perception in a dragging task [13]. Experience with a specific musical instrument might also affect the perception of audio latencies when playing it [21]. Moreover, *age* has been suggested as a factor affecting latency perception, with younger users perceiving smaller latencies than older users [21]. Closely connected to task complexity is *cognitive load*. The higher the task demands (e.g., because of higher task complexity, secondary tasks or environmental variables), the higher the user's cognitive load. This factor has been discussed with respect to latency perception in several studies [2, 19, 25].

6 Conclusion and Implications

To conclude, while several design guidelines recommend a maximum latency of 100 ms for an optimal user experience in basic interactions, empirical results suggest that latency thresholds for different tasks lay substantially lower. Users are indeed able to perceive latencies down to single milliseconds in specific tasks. Moreover, performance in zero-order and more demanding second-order tasks already gets impaired by latencies between 16–60 ms. Therefore, the lower boundary of 100 ms as mentioned in several design guidelines appears outdated. Especially interactions that are very similar to physical interactions require substantially smaller maximum acceptable latencies. Furthermore, several factors affect latency perception and consequently user performance and tolerance. Hence, a need for updated, evidence-based latency guidelines incorporating system-, task-, and person characteristics emerges.

The literature review revealed further implications. First, the majority of tasks that were utilized in empirical investigations on latency perception were zero-order tasks. However, latency can also impair user performance and experience in first- and second-order tasks. Especially in the emerging field of human-robot-interactions, virtual environments and remote-controlled systems, influences of latency should be further investigated in more complex tasks. Second, the study of factors affecting latency perception, user performance, and user experience needs to be intensified. Besides replicating previous studies and examining several variables more deeply (e.g., domain-specific experience, learning effects, attentional focus, motivational aspects), this also involves assessing age-diverse samples with varying usage experience of the utilized devices and highly dynamic computer games. Moreover, with technical progress aiming at increasingly reducing latencies, users likely get accustomed to hardly perceivable delays. This could lead to a higher sensitivity for very short latencies in users with much experience

with such modern systems and is probably one factor why guidelines from the 20th century are not applicable anymore.

Updated latency guidelines that give specific recommendations for different user groups and use cases will constitute a fruitful information source for interaction designers and system engineers and will enable a more precise and differentiated evaluation of the question: Is zero-latency really necessary?

References

1. Annett, M.: The fundamental issues of pen-based interaction with tablet devices. Dissertation, University of Alberta (2014)
2. Annett, M., Ng, A., Dietz, P., Bischof, W., Gupta, A.: How low should we go? Understanding the perception of latency while inking. In: Graphics Interface Conference 2014, 7–9 May, Montreal, Canada, pp. 167–174. Canadian Human-Computer Communications Society (2014)
3. Boucsein, W.: Forty years of research on system response times – what did we learn from it? In: Schlick, C.M. (ed.) *Industrial Engineering and Ergonomics*, pp. 575–593. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-01293-8_42](https://doi.org/10.1007/978-3-642-01293-8_42)
4. Brady, K., Wu, B., Sim, S.H., Enquobahrie, A., Ortiz, R., Arikatla, S.: Modeling reduced user experience caused by visual latency. In: Soares, M., Falcão, C., Ahram, T.Z. (eds.) *Advances in Ergonomics Modeling, Usability & Special Populations. Advances in Intelligent Systems and Computing*, pp. 267–277. Springer, Cham (2017). doi:[10.1007/978-3-319-41685-4_24](https://doi.org/10.1007/978-3-319-41685-4_24)
5. Butler, T.W.: Computer response time and user performance. In: Janda, A. (ed.) *CHI 1983, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 58–62. ACM, New York (1983). doi:[10.1145/800045.801581](https://doi.org/10.1145/800045.801581)
6. Card, S.K., Moran, T.P., Newell, A.: *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, Hillsdale (1983)
7. Card, S.K., Robertson, G.G., Mackinlay, J.D.: The information visualizer, an information workspace. In: *CHI 1991, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 181–186. ACM, New York (1991). doi:[10.1145/108844.108874](https://doi.org/10.1145/108844.108874)
8. Csikszentmihalyi, M.: *Flow and the Foundations of Positive Psychology*. Springer, Dordrecht (2014). doi:[10.1007/978-94-017-9088-8](https://doi.org/10.1007/978-94-017-9088-8)
9. Dabrowski, J., Munson, E.V.: 40 Years of searching for the best computer system response time. *Interact. Comput.* **23**, 555–564 (2011). doi:[10.1016/j.intcom.2011.05.008](https://doi.org/10.1016/j.intcom.2011.05.008)
10. Deber, J., Jota, R., Forlines, C., Wigdor, D.: How much faster is fast enough? User perception of latency & latency improvements in direct and indirect touch. In: *CHI 2015, April 18–23, 2015, Seoul, Republic of Korea*, pp. 1827–1836. ACM, New York (2015). doi:[10.1145/2702123.2702300](https://doi.org/10.1145/2702123.2702300)
11. Doherty, R.A., Sorenson, P.: Keeping users in the flow: mapping system responsiveness with user experience. *Proc Man* **3**, 4384–4391 (2015). doi:[10.1016/j.promfg.2015.07.436](https://doi.org/10.1016/j.promfg.2015.07.436)
12. Fischer, A.R.H., Blommaert, F.J.J., Midden, C.J.H.: Monitoring and evaluation of time delay. *Int. J. Hum.-Comput. Int.* **19**, 163–180 (2005). doi:[10.1207/s15327590jhc1902_1](https://doi.org/10.1207/s15327590jhc1902_1)
13. Forch, V., Franke, T., Rauh, N., Krems, J.F.: Are 100 milliseconds fast enough? Characterizing latency perception thresholds in mouse-based interaction. Paper presented at the 19th International Conference on Human-Computer Interaction, Vancouver, Canada, 9–14 July 2017 (2017)
14. Friston, S., Karlström, P., Steed, A.: The effects of low latency on pointing and steering tasks. *IEEE Trans. Vis. Comput. Graph.* **22**, 1605–1615 (2015). doi:[10.1109/TVCG.2015.2446467](https://doi.org/10.1109/TVCG.2015.2446467)

15. Goodman, T.J., Spence, R.: The effect of computer system response time on interactive computer aided problem solving. *ACM SIGGRAPH Comput. Graph.* **12**, 100–104 (1978). doi:[10.1145/965139.807378](https://doi.org/10.1145/965139.807378)
16. Ivkovic, Z., Stavness, I., Gutwin, C., Sutcliffe, S.: Quantifying and mitigating the negative effects of local latencies on aiming in 3D shooter games. In: *CHI 2015*, April 18–23, 2015, Seoul, Republic of Korea, pp. 135–144. ACM, New York (2015). doi:[10.1145/2702123.2702432](https://doi.org/10.1145/2702123.2702432)
17. Jagacinski, R.J., Flach, J.M.: *Control Theory for Humans*. Lawrence Erlbaum, Mahwah (2003)
18. Jota, R., Ng, A., Dietz, P., Wigdor, D.: How fast is fast enough? a study of the effects of latency in direct-touch pointing tasks. In: *Proceedings of CHI 2013 Conference on Human Factors in Computing*, April 27–May 2, 2013, Paris, France, pp. 2291–2300. ACM, New York (2013). doi:[10.1145/2470654.2481317](https://doi.org/10.1145/2470654.2481317)
19. Kaaresoja, T., Brewster, S., Lantz, V.: Towards the temporally perfect virtual button: touch-feedback simultaneity and perceived quality in mobile touchscreen press interactions. *ACM Trans. Appl. Percept.* **11**, 9:1–9:25 (2014). doi:[10.1145/2611387](https://doi.org/10.1145/2611387)
20. Kaaresoja, T.: *Latency guidelines for touchscreen virtual button feedback*. Dissertation, University of Glasgow (2016)
21. Mäki-Patola, T., Hämäläinen, P.: Latency tolerance for gesture controlled continuous sound instrument without tactile feedback. In: *International Computer Music Conference Proceedings*, 2004 (2004)
22. Martens, J., Franke, T., Rauh, N., Krems, J.F.: Effects of low-range latency on performance and perception in a virtual, unstable second-order control task (2016). Manuscript submitted for publication
23. Miller, R.B.: Response time in man-computer conversational transactions. In: *Proceedings of the AFIPS 1968*, December 9–11, 1968, pp. 267–277. ACM, New York (1968). doi:[10.1145/1476589.1476628](https://doi.org/10.1145/1476589.1476628)
24. Nielsen, J.: *Usability Engineering*. Academic Press, San Diego (2003)
25. Ng, A., Annett, M., Dietz, P., Gupta, A., Bischof, W.F.: In the blink of an eye: investigating latency perception during stylus interaction. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1103–1112. ACM, New York (2014). doi:[10.1145/2556288.2557037](https://doi.org/10.1145/2556288.2557037)
26. Ng, A., Dietz, P.H.: The effects of latency and motion blur on touch screen user experience. *J. SID* **22**, 449–456 (2015). doi:[10.1002/jsid.243](https://doi.org/10.1002/jsid.243)
27. Ng, A., Lepinski, J., Wigdor, D., Sanders, S., Dietz, P.: Designing for low-latency direct-touch input. In: *UIST 2012*, October 7–10, 2012, Cambridge, Massachusetts, USA, pp. 453–464 (2012). doi:[10.1145/2380116.2380174](https://doi.org/10.1145/2380116.2380174)
28. Nordahl, R.: Self-induced footsteps sounds in virtual reality: latency, recognition, quality and presence. In: *The 8th Annual International Workshop on Presence, PRESENCE 2005, Conference Proceedings*, 21–23 September 2005, London, United Kingdom, pp. 353–355 (2005)
29. Seow, S.C.: *Designing and Engineering Time: The Psychology of Time Perception in Software*. Addison-Wesley Professional, Indianapolis (2008)
30. Shneiderman, B.: Response time and display rate in human performance with computers. *Comput. Surv.* **16**, 265–285 (1984). doi:[10.1145/2514.2517](https://doi.org/10.1145/2514.2517)
31. Shneiderman, B., Plaisant, C.: *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley Publ. Co., Reading (1987)
32. Tolia, N., Andersen, D.G., Satyanarayanan, M.: Quantifying interactive user experience on thin clients. *Computer* **39**(3), 46–52 (2006). doi:[10.1109/MC.2006.101](https://doi.org/10.1109/MC.2006.101)

33. Weinberg, S.: Learning effectiveness: the impact of response time. *ACM SIGSOC Bull.* **13**, 140 (1981). doi:[10.1145/1015579.810983](https://doi.org/10.1145/1015579.810983)
34. Wickens, C.D., Hollands, J.G., Banbury, S., Parasuraman, R.: *Engineering Psychology and Human Performance*. Routledge, Oxford (2013)
35. Zhou, R., Shao, S., Li, W., Zhou, L.: How to define the user's tolerance of response time in using mobile applications. In: *2016 IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 281–285. (2016). doi:[10.1109/IEEM.2016.7797881](https://doi.org/10.1109/IEEM.2016.7797881)

Evaluation of Interface Modality for Control of Multiple Unmanned Vehicles

Gloria L. Calhoun^{1(✉)}, Heath A. Ruff², Kyle J. Behymer^{2(✉)},
and Clayton D. Rothwell²

¹ Air Force Research Laboratory, 711 HPW/RHCI, Dayton, OH, USA
gloria.calhoun@us.af.mil

² Infoscitex, Dayton, OH, USA
{heath.ruff.ctr, kyle.behymer.1.ctr,
clayton.rothwell.ctr}@us.af.mil

Abstract. The U.S. Air Force envisions future applications in which a single human operator manages multiple heterogeneous unmanned vehicles (UVs). To support this vision, a range of play-based interfaces were designed by which an operator can team with autonomy (consisting of several intelligent agents/services) to manage twelve air, ground, and sea surface UVs performing security defense tasks for a simulated military base. To enable flexible delegation control, the interfaces were designed to enable the operator to use one or more of three control modalities in calling and editing plays that define UV actions. Specifically, each step defining a play could be completed: (1) manually, via mouse/click inputs, (2) by touching a touchscreen monitor, or (3) via speech commands. This paper reports results relevant to input modality from two experiments where operators were free to choose which modality to employ. Operators overwhelmingly used the mouse compared to the touchscreen or speech and were faster and more accurate with the mouse. Subjective data also favored the mouse modality with operators commenting that it was more intuitive to use with the play calling interfaces. Results are discussed and recommendations for further multimodal research are provided.

Keywords: Multimodal interfaces · Unmanned systems · Autonomous vehicles · Speech recognition · Touchscreen

1 Introduction

Multimodal interfaces have the potential to enhance human-autonomy interaction. For example, system inputs made with speech, gesture, and touch leverage natural human communication capabilities. Thus, multimodal interface concepts should be evaluated in conjunction with other technology advancements towards enabling single operator management of multiple heterogeneous unmanned vehicles (UVs). The Air Force

The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.

Research Laboratory (AFRL) recently led a multi-service effort to integrate several autonomy advancements into a control station prototype referred to as “IMPACT” (Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies) [1]. The interfaces in the IMPACT system were designed to support a wide spectrum of human-autonomy control of multiple UVs (air, ground, and sea surface) as they perform dynamic security mission tasks defending a simulated military base [2]. At one extreme, the operator calls “plays” that define the actions of one or more UVs. With this play-based adaptable automation approach, the operator can quickly task UVs by specifying high level commands indicating the play type and location, and the autonomy determines all other parameters. For example, when an IMPACT operator calls a play to achieve air surveillance (play type) on a building (location), an intelligent agent recommends a UV to use (based on estimated time en route, fuel use, environmental conditions, etc.) and a cooperative control algorithm provides the quickest route to get to the building (taking into account no-fly zones, etc.). At the other end of the control spectrum, the operator can manually control UV movement with keyboard/mouse inputs or build plays from the ground up with minimal autonomy assistance. Between these two extremes of the control spectrum, the operator makes more inputs to the play, for instance specifying parameters and constraints that the autonomy may not be aware of (e.g., current visibility which can drive UV/sensor payload choice).

Besides providing the operator flexibility on the degree to which the autonomy assists with UV control, the interfaces in IMPACT were also designed to provide the operator flexibility in terms of which control modality could be employed to make inputs [1]. Specifically, plays could be called or edited: (1) via mouse/click inputs, (2) by touching a touchscreen monitor, or (3) via speech commands. These multimodal inputs support the overarching architecture that allows the operator to flexibly interact with autonomy at any time, employing any of the three control modalities. In other words, the interfaces were designed to support all three modalities for each step in utilizing the play-based interfaces. This approach was based on past research that has shown that most users prefer interfaces that are multimodal versus unimodal (e.g., 56–89% of users in an evaluation comparing spoken, written, and combined pen/speech input [3]). This preference reflects a number of advantages of having multiple control modalities available. First, the operator may prefer selecting which modality to employ, even alternating between input modalities to capitalize on the advantages of each modality [4]. Having multiple modalities available helps prevent the overuse of any individual mode. Also, some modalities may be more aligned with certain types of tasks or environmental situations [5]. For instance, making inputs with speech commands may not be ideal if the operator is involved in conversations or is in a noisy environment. Certain types of information are less amenable to vocal specification too (e.g., temporal relationships) [6]. Having multiple modalities available also allows the operator to leverage knowledge and past experience with respect to when and how to deploy a modality for the most efficient and accurate inputs [4].

Past research supports the use of multimodal interaction in UV applications. The utility of touch [7], speech-based input [8, 9], and “spatial dialog” (a combination of speech and touch input [10]) has been examined for single air UV control. In multi-UV control research by Levulis and colleagues [11], participants supervised a team of three air UVs and two manned helicopters traveling towards a landing zone to deploy ground

troops. Results showed that both touch and multimodal (touch and speech) input conditions were better than the speech-only condition in terms of task performance and subjective ratings of workload, situation awareness, and input usability. Their tasks focused on the input modality for monitoring and reporting status (e.g., classifying photographs, responding to instrument warnings, and addressing task queries), in contrast to the present research that emphasizes play-based UV control.

For play-based interfaces that establish respective human/autonomy roles in task completion, prior research has also shown the benefits of multimodal input. In a simulation demonstration of multiple input methods (keyboard/mouse, touch, and speech) for calling *single* air UV plays, pilots commented that the use of speech input was a natural method, but multimodal options should be available since certain tasks lend themselves to one mode versus another [12]. There were, however, concerns about the vocabulary training and memory requirements if the number of plausible plays and associated parameters is large. Additionally, speech commands should have a meaningful relationship (e.g., semantic) to their resulting actions [13].

Multiple air UVs were successfully controlled via plays called with speech recognition in a flight demonstration of a delegation control interface used in an urban mission scenario. Mean reaction time to mission events was significantly shortened with speech recognition, reflecting the ability for the operator to bypass cumbersome menu control steps [14]. This was viewed as especially advantageous during time critical mission phases. Similar advantages for play-based speech control were found in a simulation evaluation comparing multimodal inputs for control of three air UVs [15, 16]. Participants could either call plays with speech or by performing drag/drop actions to move symbology with the mouse or finger into “activity windows” used to construct plays for one or more UVs. Data on which control modality was used most frequently was not reported. However, participants commented that even though the ability to employ multiple input methods was useful, input would be even more flexible if the operator could switch between modalities while calling a single play. In other words, all modalities should be available in specifying a play such that the operator can flexibly switch between methods on a step-by-step basis [17].

The present paper will describe the multimodal play-based control approach used in two recent experiments employing the IMPACT simulation (Fig. 1).



Fig. 1. IMPACT simulation

Across the experiments, data were collected from fourteen participants familiar with unmanned vehicle operations and/or base defense missions. Operators received briefing and training, including multimodal control practice, for base defense mission related tasks involving simulated air, ground, and sea surface UVs. Each of the two experiments will be described separately, followed by a summary discussion. For each experiment, an overview of the play-based interfaces and methodology will be provided. This will be followed by report of the results that specifically pertain to which modality (mouse/click inputs, touch, or speech) was employed when interacting with the play-based interfaces, as well as other modality relevant objective and subjective data. However, first a brief overview of key elements of the play-based interfaces will be provided as well as methodology details common across the two experiments.

2 IMPACT Play-Based Interface Approach

Given that the previous research [12–17] supported relatively few plays (and primarily a single UV type), extensive development was required to implement play-based interfaces in IMPACT for heterogeneous UV control (see [2, 18]). The design and implementation process was incremental. Experiment 1 provided only a few play-based interfaces to control 6 UVs. In contrast, Experiment 2 featured refinements for the interfaces utilized in Experiment 1, as well as additional interfaces to better support control of 12 UVs. The differences in play-related interfaces between the experiments will be described in Sects. 3 and 4. As an introduction, this section provides an overview of common elements (see [2, 18] for more details).

2.1 IMPACT Simulation

In both experiments, operators sat at the IMPACT control station supported by an AFRL developed Fusion software framework for coordinating the communications from multiple systems and software components (for more details, see [19]). The control station was designed to support simulated single operator management of multiple heterogeneous UVs performing a base defense mission. (The mission details and tasking were informed by an earlier cognitive task analysis [20].) The operator's key task was to respond to mission events that required the reassignment of one or more UVs from normal patrol to instead either investigate a threat or perform other defensive measures (e.g., surveil the ammo dump every 30 min).

The control station for both experiments contained a keyboard, mouse, foot-pedal (for push-to-talk speech control), and a Plantronics GameCom Commander headset with boom microphone (for speech input and audio feedback). (Experiment 2 contained an additional foot-pedal for radio communication with a confederate sensor operator.) Both experiments employed four monitors (Fig. 1). The top center monitor presented a Tactical Situation Display (TSD) that included a geo-referenced map showing the locations of each UV and its associated on-going patrol route (white symbology, with gray UV symbols), route if under manual control (dark gray), or ongoing play. Symbology for each different play was presented in a unique color; if the play involved multiple

UVs, all the UVs and respective routes were presented in the same color. Lines depicting routes were coded to differentiate ongoing patrol and plays (solid lines) from plays being developed (dashed lines).

Both the left and right monitors were considered auxiliary displays (Fig. 1). The left monitor presented “help” information related to the mission and the right monitor presented imagery from each UV’s sensor (simulated via SubrScene Image Generator; www.subrscene.org/). While operators were presented payload information (as well as symbology showing each sensor’s field-of-view on the map), the operators’ responsibility in both experiments was to manage the movement and tasking of the UVs. Operators were briefed that a remotely located sensor operator (simulated in Experiment 1) was tasked to monitor and interpret the sensor imagery and that assessment information would be communicated from the sensor operator and/or commander via chat (and/or radio in Experiment 2).

The lower center sandbox (touch sensitive) monitor (Fig. 1) presented elements of the TSD, as well as several interfaces pertaining to play calling and management. This monitor provided a workspace for operators to interact with the UVs and autonomy support without obscuring the current state of the world (which was always visible on the TSD). Interactions involved either keyboard/mouse/touch inputs to manually control a UV’s movement (Experiment 2 only) or play-based interfaces to call and edit single- and multi-UV plays. In both experiments, three input modalities (mouse/click, touch, and speech) were implemented for operators to interact with the play-based interfaces. All three input modalities were identical in terms of the control actions initiated, as well as other control station feedback.

2.2 IMPACT Control: Pictorial Symbology and Speech Commands

Concise symbology (illustrated in Fig. 2) was used on the sandbox’s map and play interfaces to represent each UV and play type [18]. Each UV symbol was shape coded (e.g., air: plane form, ground: wheeled rectangle, surface: finned pentagon). Each play type was depicted by a circle with inner pictorial symbology representing the UV(s)’ task (e.g., plus sign to surveil either a point/location, line for road, and square for an area). The UV(s) associated with each play was represented by both shape and location coding on the play icon’s surrounding circle (e.g., air UV upper left versus ground UV lower left). Concise symbology was also designed to represent many play-related details

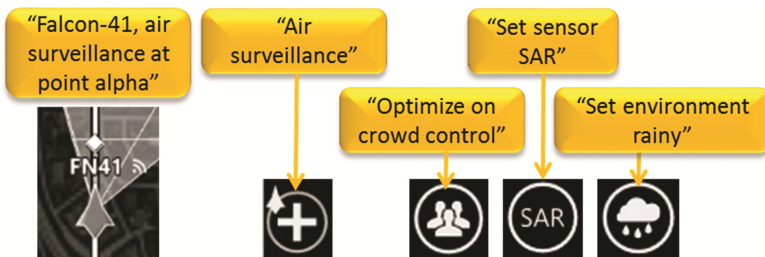


Fig. 2. Sample UV and play icons and associated speech commands

(Fig. 2) such as the target size, current environment, play priority, and what factors/constraints are pertinent to the play [18].

Each pictorial symbol utilized in the play-related interfaces presented UV/play related information and acted as a control element to initiate the play. Selecting the symbol with either a mouse click or single finger touch and release (“lift off” or “last contact” touching) changed the UV/play functioning in some manner, affording the operator the advantages of direct perception [21] and manipulation [22]. With these two manual input modes, the operator directly acted on the object of interest.

To implement the speech input mode, a companion speech command (either a word or phrase) was determined for each manual input. To illustrate, Fig. 2 provides the speech commands for a sample of icons used in the play interfaces for specifying play type and detail. During experiments, speech command reference information was displayed on the left auxiliary monitor. (In Experiment 1, the speech system had a vocabulary of 84 words and was capable of recognizing and parsing 2160 phrases. The system was expanded in Experiment 2 to 322 words with the capability of recognizing and parsing tens of millions of phrases.) In both experiments, a push-to-talk (PTT) approach was used to differentiate speech commands issued into a headset from other auditory communications. Operators signaled the Sphinx speech recognition system [23] to start processing the verbal input by either depressing a pedal on the floor or by clicking or touching a bar on the lower monitor. After the PTT switch was released, operators could confirm if the command was recognized from the auditory and visual feedback provided. (The visual feedback was presented on the PTT bar for 2 s before fading away and also displayed in a scrolling chat window exclusively dedicated to speech interaction).

2.3 IMPACT Familiarization for Experimental Participants

Experimental sessions began with operators completing a demographics questionnaire. Next, operators were given a simulation overview describing the project’s goals and introducing the concept of play calling. Operators were then seated at the IMPACT station and given a mission briefing that included:

- A description of the UVs they would be controlling, how each UV, its route, and its sensor footprint were represented on the map, and the tasks that each UV was responsible for performing in support of base defense operations.
- An overview of the base they would be defending including the base’s perimeter, sectors, critical facilities, patrol zones, and the named areas of interests in the area immediately surrounding the base.
- An explanation of their role as a multi-UV operator supporting base defense operations: in response to chat messages from a remotely located commander and sensor operator (played by confederates), they would be assigning high-level tasks to the UVs while the autonomous system components flew, drove, and operated the UVs.

The experimenter then provided the operator with a high-level description of the IMPACT interfaces. A detailed explanation of how plays could be called by using speech commands and/or clicking or touching icons designating play types and details was given, followed by providing time for operators to interact with the system until they

reported being familiar with the various interfaces and three input modalities. Additional methodological details are provided in the next two sections.

3 Experiment One

The first experiment was designed to evaluate the initial IMPACT interfaces in a 20-minute trial, as well as identify potential design improvements. Operators were provided 13 plays to task six UVs (three air, two ground, and one sea surface) in performing the base defense mission. Operators called plays using either speech commands (e.g., “air surveillance”) or making touch or mouse click inputs on a Play Creator interface (Fig. 3a). Once a play was initiated, operators selected a play location either with a speech command (e.g., “at the flight line”), from a dropdown menu of previously identified points, or by clicking on the map (see Fig. 3b). Further play details could be specified with speech commands and mouse/touch inputs by expanding the interface to show additional views of play-related options (Fig. 4). The operator could also view a list of active plays and the progress of plays related to several parameters. For more detailed descriptions of play calling and monitoring interfaces, as well as methodology and evaluation results, see [20].

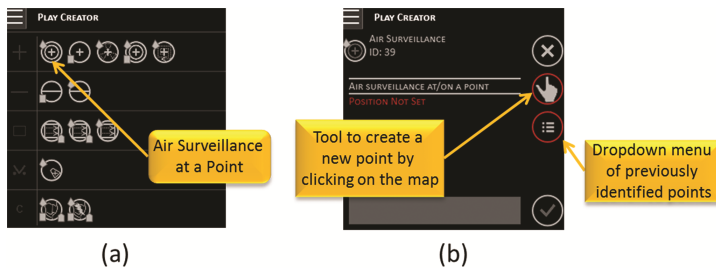


Fig. 3. Play Creator interface. (a) icons/buttons for each of Experiment 1’s thirteen plays. (b) methods for identifying play location [20]

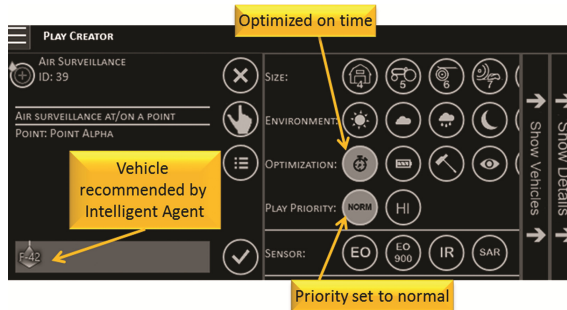


Fig. 4. Play Creator interface: Sample mechanisms for autonomy and operator to communicate UV and other constraints and details used in generating play plans [20]

3.1 Method

Participants. Seven volunteers from a U.S. Air Force Base participated. Three operators had prior experience flying UAVs (Predator, ScanEagle, Global Hawk, Shadow) as well as manned aircraft. Four operators were active Air Force security force personnel with experience conducting base defense operations in deployed environments (Afghanistan, Germany, Iraq, Kuwait, and Saudi Arabia). All operators were male and reported normal or corrected-to-normal vision and normal hearing.

Equipment. Six computers were used for IMPACT in Experiment 1 (a Dell T5610 & five Dell R7610 s running Windows 8.1). One computer ran IMPACT and the AMASE (AVTAS: Aerospace Vehicle Technology Assessment and Simulation - Multi-Agent Simulation Environment) vehicle simulation (used to simulate the UVs). One computer ran the test operator console and simulation for simulated entities in the sensor videos (Vigilant Spirit Simulation [25]), three computers ran two simulated (SubrScene) sensor videos, and one computer ran an XMPP (Extensible Messaging and Presence Protocol) Chat server for simulated communications. This IMPACT version used four 68.58 cm touchscreen monitors (Acer T272HUL; usable touch screen area: 59.69 × 33.66 cm; 2560 × 1440 resolution; tilted 45° from horizontal).

Procedure. After the general overview of the IMPACT simulation, mission-related tasks, and input modalities available for play calling, operators received a detailed briefing on the play-related interfaces available in Experiment 1. Next, training focused on providing operators with experience with each input modality. Operators received 12 chat messages asking them to call a play using a specific modality (e.g., “Using speech, call an air surveillance at Point Alpha”). Table 1 lists the exact sequence of twelve plays operators were asked to call during this portion of the training as well as the modality operators were instructed to use (4 plays for each modality).

Table 1. Play calling modality training

Method	Play
Speech	Air surveillance at point alpha
Speech	Normal full coverage patrol
Speech	Ground inspect at route brave
Speech	Air expanding square at point charlie
Touch	Air sector search at point alpha
Touch	Surface parallel search at area delta
Touch	Air parallel search at area bravo
Touch	Ground inspect at point charlie
Mouse	Air ground surveillance at point bravo
Mouse	Air inspect at route charlie
Mouse	Air surface parallel search at area delta
Mouse	Go highly mobile at point alpha

Operators were then trained on how to specify constraints, vehicles, and details when calling and/or editing a play. For all three input modalities, operators were asked via chat messages to call a specific play (e.g., “Using speech, call an air surveillance on Point Alpha, set sensor to EO, and optimize for low impact”), then make edits to the ongoing play (e.g., “Change the loiter type to a figure 8”). If an operator made a mistake, the experimenter provided feedback and the operator tried again until he had successfully completed the correct action. On average, training lasted one hour and was followed by a short break before the experimental scenario.

The goal of the 20-minute experimental scenario was to provide operators with the opportunity to exercise all of IMPACT’s capabilities within a realistic base defense scenario. Operators were informed that the scenario would begin with an air UV investigating a suspicious watercraft with all other UVs on a high alert patrol. Table 2 lists the exact sequence of mission events that occurred. Operators were instructed to respond to each chat message by calling one or more plays that best addressed the event. Operators were free to choose which of the three input modalities to employ in completing each step of the play calling process.

Table 2. Sequence of Mission Events

Event	Description
1	Operator receives a chat message from the sensor operator that the unidentified watercraft is a fishing boat
2	Operator receives a chat message from the commander to resume normal base defense operations/patrols.
3	Intelligent agent recognizes a serendipitous surveillance opportunity (an air UV is near a critical facility) and recommends a play (air surveillance at the critical facility) to the operator.
4	Operator receives a chat message from the commander to send an air UV to point charlie and to instruct other UVs to go highly mobile in response to a patrol reporting smoke at point charlie.
5	Operator receives a chat message from the sensor operator to send a ground UV to point charlie to get a better look.
6	Operator receives a chat message from the commander to provide the ground UV headed to point charlie with an air UV overwatch.

Once the experimental scenario was completed (approximately five plays called), operators completed paper questionnaires on the overall IMPACT system and its components. Then a semi-structured interview was conducted to capture additional feedback on IMPACT and its associated technologies including the three different input modalities. The entire procedure lasted approximately 3.25 h.

3.2 Results

Of the seven individuals who participated in the study, six completed the training and mission in the allotted time. Due to unanticipated time restrictions, the seventh operator was unable to complete the study and was eliminated from the data analysis. Due to the

small number of participants, UV operators and security force personnel data were not analyzed separately.

Subjective feedback on touch and speech was mixed. In general, operators seemed to like the idea of being able to execute plays via touch and speech. However, operators expressed concerns about the touchscreen’s lack of precision (an operator might touch an icon three times before the system registered it) and the speech system’s poor accuracy (an operator might utter the same command several times before it was recognized; the word error rate was 21.95% for in-grammar utterances).

Objective data were collected on the modality (mouse, touch, or speech) that operators used to call plays during the experimental mission (when operators choose which input modality to employ). Operators used the mouse much more frequently than touch or speech (see Fig. 5). Note that data for ‘speech’ is labeled “Speech/Mouse Confirm” because when operators used speech during the mission, they always used it in conjunction with the mouse: operators would initiate a play call with a speech command but execute the play by clicking the checkmark with the mouse instead of saying the speech command “Confirm” to execute the play. Actually, only two operators employed the speech modality during the experimental trial, and a different operator (only one) employed the touch input modality. Operators made a higher percentage of major errors (defined as failing to complete a play) when using touch than mouse or speech (see Fig. 5).

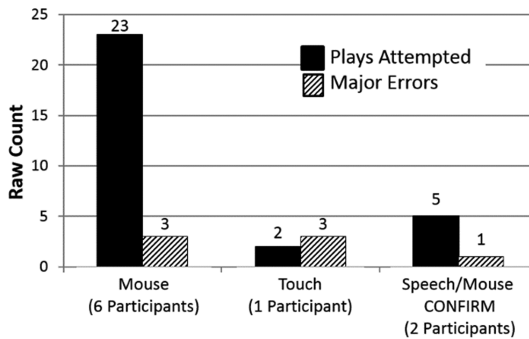


Fig. 5. Number of plays attempted and number of errors by input modality

Operators were also faster at completing plays using the mouse as compared to using touch or speech (Fig. 6a). However, this difference most likely reflected the reported problems operators had with touch input. In fact, when only plays correctly completed (i.e., no major errors) were examined, the mean difference between time to complete a play with the mouse and speech was just 2.5 s (see Fig. 6b). Note that operators never correctly completed a play using touch input.

Operators overwhelmingly used the mouse input modality compared to the touch or speech, and were faster and more accurate with the mouse as well. Several factors may have contributed to these results. Multiple operators had difficulties with the touchscreen registering inputs, but commented that if the touchscreen had worked better they would have been more likely to use it. For example, one operator stated, “Touchscreen could be extremely intuitive and quick if implemented correctly.”

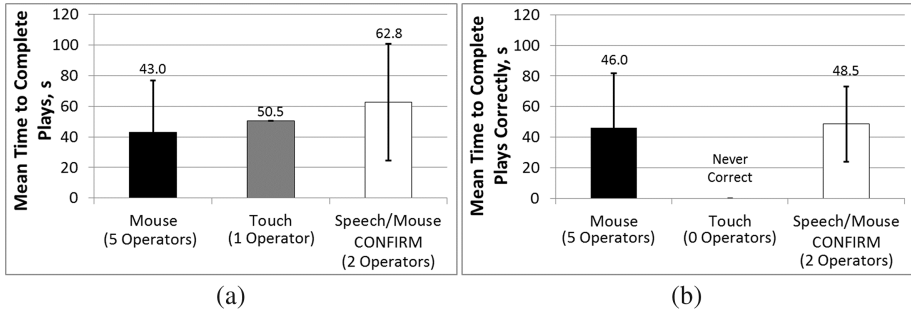


Fig. 6. Mean time to complete a play call by modality: (a) all plays, (b) plays called correctly. (Error Bars are the standard deviations.)

Several operators also spoke favorably of the speech input modality, especially the security force personnel, who mentioned that the speech commands were very similar to the dispatch calls they make during security force operations. However, this preference was not reflected in performance, as operators used the mouse more than speech to call plays. Several operators commented that they weren't completely familiar with the speech vocabulary, suggesting inadequate training. In the end, operators may have chosen to use the mouse modality due to its reliability; clicking a play icon with the mouse consistently resulted in the desired action, while touching a play icon or issuing a speech command often failed to register an input.

3.3 IMPACT Modifications

Based on the results of Experiment 1 (see Sect. 3.2), several modifications were made to improve IMPACT's touch and speech input modalities. For touch input, the main concern was that the sizes of selectable areas were too small (e.g., for play icons in Fig. 3a: 6.35 mm diameter circles with 1.59 mm separation). Although smaller targets (1.7 mm) were selectable in earlier research [26], MIL-STD-1472G [27] recommends a 15.2×15.2 mm area. To help aid play icon selection in Experiment 2, the diameter of the play icon's selectable area was increased slightly (7.94 mm diameter). Additionally, the touchscreen was replaced with a slightly larger one positioned at a lower tilt angle (see Sects. 3.1 and 4.1).

For the speech modality, the finite grammar was dramatically expanded to allow hundreds more ways to say things, resulting in a large increase in flexibility and naturalness. Commands were also added to support a more complex mission (i.e., more UVs, larger variety of play types, and ability to specify play details with speech). By modifying the speech pipeline, the operator in Experiment 2 could change symbology clutter level with speech and issue verbal queries to autonomy (e.g., "which vehicle can get to the flight line the fastest?" followed by aural and text responses). This process attempted to strike a balance between flexibility and accuracy, as changing from a closed- to an open-language vocabulary can increase recognition errors dramatically. Besides this expansion of the speech system language model, the acoustic model was changed, based on extensive testing.

4 Experiment Two

Besides the modifications described in Sect. 3.3, the simulation and play-based interfaces were expanded to provide operators 25 base defense related plays, in addition to two types of patrols. The assets were also increased to 12 UVs (four each of air, ground, and sea surface), with more variety in payload, including some with weapons. This involved making many changes to the symbology and interfaces used in Experiment 1, besides the modifications prompted by the operators’ comments. Figure 7 illustrates the revisions made to the interfaces used in Experiment 1 to call a play and support operator-autonomy communications on play details. For more details on the interfaces for play calling and monitoring, see [24].



Fig. 7. Play Calling and Play Workbook used in Experiment 2 [24]

In addition to revising the interfaces used in Experiment 1, additional play based interfaces were added. Two of them provided other means to employ mouse and touch input modalities. One interface termed “radial menu” allowed operators to call plays directly from the map, instead of utilizing the play-calling interface illustrated in Fig. 7a. When the operators selected a location on the map (Fig. 8a) or a UV on the map (Fig. 8b), a radial menu appeared consisting of only the play options relevant to that location or UV (e.g., no ground based plays if a sea surface UV selected). The radial menu appeared with a right click of the mouse or if the operator touched the screen with a finger, continued the touch (i.e. lingered), and released upon feedback (white square presented) that the selection was registered.

The second additional play-calling interface was available in a task manager that maintained a list of tasks to be completed based on prompts in the chat window and previously defined (quick reaction checklist) steps for addressing certain mission events. As shown in the illustration (Fig. 9), the top row contains the mission prompt from chat (“Unidentified Watercraft...”) and each row below it shows established tasks that should be performed in response to such an event. Selecting each of these rows (a single touch and release, signaling left mouse click) called up a row below it that either displayed more detailed task text or a play button to select and call up the corresponding Play Workbook (Fig. 7b) to further support the play calling.

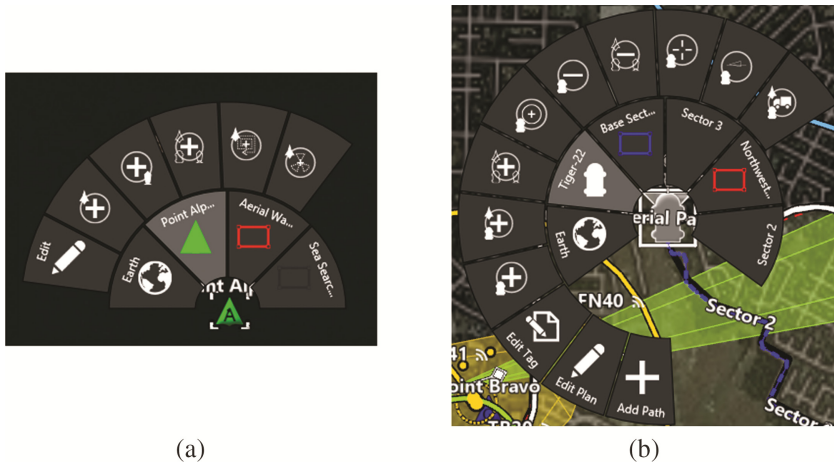


Fig. 8. Illustration of radial menu when: (a) UV selected or (b) location selected

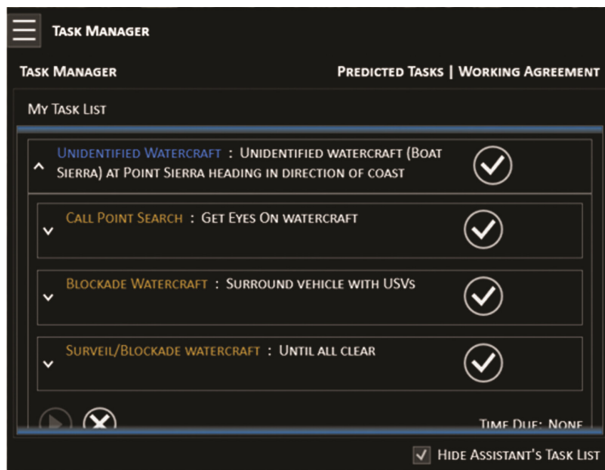


Fig. 9. Illustration of task manager that provided a mechanism for calling plays with mouse and touch input modalities

Experiment 2 also featured added interfaces that provided operators with further insight into the rationale of the plans generated by the autonomy for plays as well as the status of plays under development. These interfaces primarily presented information to the operator to monitor play calling and execution and are of less interest to this input modality-focused discussion. (For further details, see [2, 18, 24].)

In Experiment 2, the IMPACT prototype was compared to a baseline system (representing current state-of-the-art [25] that did not have play-based interfaces, including speech control. Four 60-minute experimental trials were conducted, two trials with each

system with the trials varying in mission complexity (number and timing of mission-related tasks). Trials were blocked by system and counterbalanced. Only data from the two trials conducted with the IMPACT system were relevant to exploring operators' modality choice when making inputs into the play-based interfaces.

4.1 Method

Participants. Eight volunteers with relevant military experience participated in this study, four active duty and four who had previously served. Six operators had prior experience piloting air UVs (Global Hawk, Predator, Reaper, Scan Eagle, Raven), one operator was a former Predator/Reaper SO, and one operator was an experienced security force and base defense expert. Seven operators were male (one female) and all operators reported normal or corrected-to-normal vision, normal color vision, and normal hearing. Operators' mean age was 43.6 years (SD = 10.84).

Equipment. The experimental configuration was expanded to four stations: the C2 Operator Station, the Sensor Operator Station, the Test Operator Console, and the Simulation Station. The Simulation Station used a Dell Precision T5400 and ran One Semi-automated Forces (OneSAF), a simulation tool that generated all friendly, neutral, unknown, and hostile forces during the experiment, with the exception of the UVs. The C2 Operator Station and Test Operator Console each used a Dell Precision T7910 while the Sensor Operator Station used a Dell Precision T5600. The C2 Operator Station, Sensor Operator Station, and Test Operator Console had identical monitor setups, with three Acer T272HUL LED touchscreens (2560 × 1440) and one (lower center) Sharp PN-K322B 4 K Ultra-HD LCD touchscreen (usable touchscreen area: 69.79 × 39.26 cm; 3840 × 2160 resolution; tilted 42° from horizontal). Three Dell Precision R7610 located in a different room provided the sensor feeds for the UVs (four feeds per machine).

Procedure. Operators received briefings and training, as described in Sect. 2.3, as well as training on the play-based interfaces added in Experiment 2. Practice trials were also conducted to provide operators familiarity on how mission events would be prompted (chat window and over the headset) and how to respond in trials featuring IMPACT play-based interfaces versus the baseline system (the latter not described here). In addition to how to respond to mission events, operators were familiarized with how to accomplish anti-terrorism measures assigned at the beginning of each trial (e.g., image four sides ("360") of a certain building at a set interval (accomplished by calling point inspect plays in a timely manner)). There were also other base defense tasks added in Experiment 2 (such as to provide a ground vehicle scout ahead coverage with an air UV), in addition to queries issued by the commander through chat (e.g., "How long would it take to get a Show of Force at Gate 3 in place?"). For most queries, it was more efficient for operators to issue speech commands to the autonomy, asking for relevant information to answer the query. Training to address all these tasks took an entire day to accomplish. On a second day, after refresher training, the four experimental trials were conducted and several debriefing questionnaires were administered.

4.2 Results

Operators' performance was better on multiple mission performance metrics with the IMPACT system as compared to the baseline system. For instance, operators were able to execute plays using significantly fewer mouse clicks with IMPACT compared to baseline. Operators also rated IMPACT higher on usability than the baseline. Detailed results will be published elsewhere [24]. Here, results will focus on operators' use of the three input modalities (mouse/click, touch, and speech).

In that Experiment 2's focus was on evaluating the IMPACT and baseline systems, operators were not asked to compare the modalities in questionnaires. Figure 10 illustrates this point by showing data from three Likert Scales addressing IMPACT features with respect to: how easy to use, how quick to learn, and potential value for future multi-UV operations. (Likert scales were five-point, ranging from 'Strongly Disagree' to 'Strongly Agree.') The administered questionnaire specifically had items addressing touch and speech input modalities. However, there was not a specific scale for mouse input. In Fig. 10, data for the mouse input modality are estimated from the scale for the Play Calling interface (Fig. 7a), as the experimenter observed that none of the operators used touch to interact with this interface. It cannot be determined, though, if the operators' ratings reflect the mouse modality and/or other features of the play calling interface (e.g., arrangement of play icons in the rows).

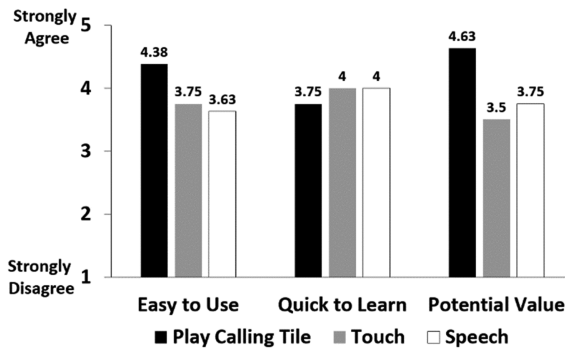


Fig. 10. Ratings related to each input modality

Four operators reported that there were difficulties making inputs with touch (e.g., citing lack of confidence and its sensitivity to "fat fingers"). One operator reported that after employing the mouse input, it didn't make sense to change to another modality. Comments pertaining to the speech input (from three operators) mentioned its unreliability and restrictive syntax. The word error rate was 23.38% for Experiment 2 (for in-grammar utterances). This error rate was only slightly better (~1.5%) from that achieved in Experiment 1, despite increases in both vocabulary and phrases. The word error rate across all utterances, both in-grammar and out-of-grammar, was 34.26%.

The degree to which operators did not employ touch and speech input is further illustrated in Fig. 11 that shows the number of plays called with each type of play interface, as well as the modality employed (given all interactions with the radial menu, play

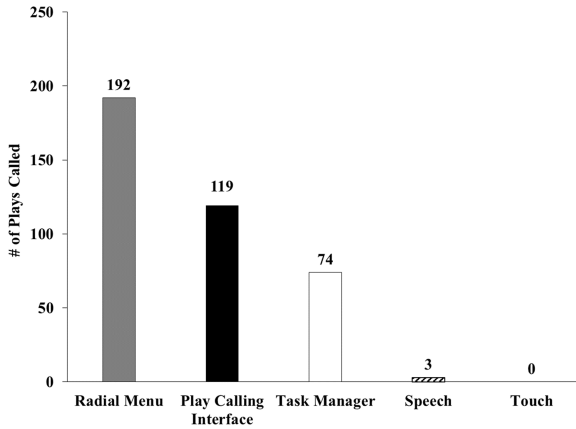


Fig. 11. Number of plays called with each play-based interface and input modality type

tile, and task manager were made with the mouse). Of the 388 play calls made during Experiment 2, only three used speech commands and none were made using touch input. These data also suggest that operators did not switch between modalities in completing steps in the play calling process. In fact, the tendency to use the mouse to confirm the speech input reported in Experiment 1 was not observed in Experiment 2. Rather, two operators who employed speech failed to make the confirmation response immediately after calling the play verbally (leaving the Play Workbook open for that play while calling other plays and then eventually closing the workbook for the verbally-called play), escalating the play calling completion time (mean = 14.23 min).

Results depicted in Fig. 12 show that every operator primarily used the mouse input modality. Also, seven of the eight operators employed the radial menu. However, the results also suggest the value of providing a variety of interfaces for calling plays as

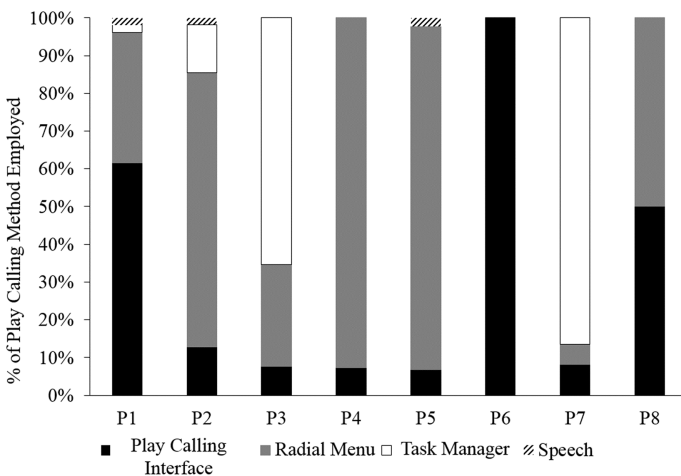


Fig. 12. Percentage of plays called with each play-based interface

three operators used the task manager to call plays (the primary mechanism for two of the operators) and three other operators primarily used the play-calling interface.

5 Discussion

Subjective data from the two experiments indicated that the mouse input modality was better than both touch and speech modalities for play-based management of multiple UVs. Operators preferred the mouse input modality for calling plays, commenting that the mouse was more intuitive for exercising the play calling interfaces. In both experiments, this preference was also evident in the frequency operators chose to employ the mouse for calling plays, rather than touch or speech. Use of mouse inputs proved to be an efficient input modality in the IMPACT simulation.

The other two modalities for play calling were problematic: operators cited the speech recognition rate and the touchscreen's lack of precision. While modifications were made in the implementation of these modalities after Experiment 1, these changes may have actually limited their utility. For instance, it is possible that more speech command training was needed to reap the benefits of expanding the speech model and vocabulary. (Note: a concern for vocabulary training requirements was raised in earlier play-related research [13].) It may also be that use of speech input is simply not ideal for play calling. As Cohen and Oviatt [6] explain, speech-based control is most ideal when the operator's hands and/or eyes are busy or when there is limited screen real estate to exercise control. With the tasking and mission employed in IMPACT, operators were able to devote attention to the play-based interfaces and employ mouse inputs efficiently.

With respect to the touch input modality, the changes to the monitor's size and tilt in Experiment 2 may have complicated touch entry by increasing the reach envelope even more, exceeding distances recommended in [26]. Touch input is more useful with smaller reach distances and when inputs are not frequent [28]. Given the missions were designed to require frequent play calling, it is logical that small manipulations of the mouse to position the cursor on various play related interfaces was more effective than reaching to touch interfaces or map locations on a large monitor. Over the course of hour-long missions (Experiment 2), issues of fatigue noted in other examinations of touch input (e.g., [28]) would have likely been observed. Operators may have also been hesitant to employ touch due to arm/hand movement occluding map/play symbology on the monitor.

These research results suggest that providing multiple input modalities for exercising IMPACT's play interfaces was not advantageous to the operators in these experiments. It could be viewed that these results suggest input modality should instead be optimized for specific tasks, rather than expending effort to enable multiple input modalities for every task type. However, this does not mean that only a single modality should be implemented for each task type. For instance, the present results indicate that the radial menu play calling interface on the map was most frequently employed, compared to the dedicated play calling interface, task manager, or utilizing speech control (Fig. 11). If direct interactions with the map (designating a location or UV) are ideal for calling a play, perhaps an integration of sketch and speech map inputs would be useful [29],

enabling the operator to draw on the map with a companion speech command to call a play (“loiter here for 10 min) or specify a play detail (“ingress here”). Continued use of the mouse for sketching is likely more convenient compared to switching modality to touch. However, a mechanism (e.g., button press or speech commands) would be needed to signal the start and end of the sketch input to differentiate it from other mouse inputs.

In addition to multiple modalities available for play calling, the operators were also free to choose modality for responding to query prompts. Despite the poor speech recognition rate, operators’ chose the speech modality to acquire most of the information needed to address these prompts. Additionally, several operators suggested that they found it useful to query the autonomy in this manner. It is recommended that this capability be expanded, along with improvements to the speech recognition system, to better support collaboration and joint problem solving between the human operator and autonomy.

Acknowledgments. This research supports the ASD/R&E Autonomy Research Pilot Initiative “Realizing Autonomy via Intelligent Adaptive Hybrid Control.”

References

1. Draper, M., Calhoun, G., Ruff, H., Frost, E., Evans, D., Hansen, M., Douglass, S., Spriggs, S., Bearden, G., Howard, M., Behymer, K., Patzek, M., Rowe, A.: Intelligent multi-unmanned vehicle planner with adaptive collaborative/control technologies. In: International Symposium on Aviation Psychology (2017)
2. Calhoun, G., Ruff, H.A., Behymer, K.J., Mersch, E.M.: Operator-autonomy teaming interfaces to support multi-unmanned vehicle missions. In: Savage-Knepshield, P., Chen, J. (eds.) *Advances in Human Factors in Robots and Unmanned Systems. Advances in Intelligent Systems and Computing*, vol. 499, pp. 113–126. Springer International Publishing, Switzerland (2017)
3. Oviatt, S.L., Olsen, E.: Integration themes in multimodal human-computer interaction. In: Shirai, K., Furui, S., Kakehi, K. (eds.) *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, pp. 551–554. Acoustical Society of Japan (1994)
4. Oviatt, S.: Ten myths of multimodal interaction. *Commun. ACM* **42**(11), 74–81 (1999)
5. Turk, M.: Multimodal interaction: a review. *Pattern Recogn. Lett.* **36**, 189–195 (2014)
6. Cohen, P.R., Oviatt, S.L.: The role of voice input for human-machine communication. *Proc. Nat. Acad. Sci.* **92**, 9921–9927 (1995)
7. Durlach, P.J., Neumann, J.L., Bowens, L.D.: Evaluation of a touch screen-based operator control interface for training and remote operation of a simulated micro-uninhabited aerial vehicle. In: Cooke, N., Pringle, H., Pedersen, H., Connor, O. (eds.) *Human Factors of Remotely Operated Vehicles*, pp. 165–178. Elsevier, NY (2006)
8. Williamson, D.T., Draper, M.H., Calhoun, G.L., Barry, T.P.: Commercial speech recognition technology in the military domain: results of two recent research efforts. *Int. J. Speech Technol.* **8**(1), 9–16 (2005)
9. Calhoun, G.L., Draper, M.H.: Multi-sensory interfaces for remotely operated vehicles. In: Cooke, N., Pringle, H., Pedersen, H., Connor, O. (eds.) *Human Factors of Remotely Operated Vehicles*, pp. 149–163. Elsevier, NY (2006)

10. Chun, W.H., Spura, T., Alvidrez, F.C., Stiles, R.J.: Spatial dialog and unmanned aerial vehicles. In: Cooke, N., Pringle, H., Pedersen, H., Connor, O. (eds.) *Human Factors of Remotely Operated Vehicles*, pp. 193–208. Elsevier, NY (2006)
11. Levulis, S.J., Kim, S.Y., DeLucia, P.R.: Effects of touch, voice, and multimodal input on multiple-UAV monitoring during simulated manned-unmanned teaming in a military helicopter. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, p. 132 (2016)
12. Calhoun, G., Draper, M.H., Ruff, H., Barry, T., Miller, C.A., Hamell, J.: Future unmanned aerial systems control: feedback on a highly flexible operator-automation delegation interface concept. In: *AIAA Infotech @ Aerospace Conference, AIAA-2012-2549*, pp. 1–16 (2012)
13. Calhoun, G., Ruff, H., Miller, C., Murray, C., Hamell, J., Barry, T., Draper, M.: Flexible levels of execution-interface technologies (FLEX-IT) for future remotely piloted aircraft control applications. Technical report, AFRL-RH-WP-TR-2012-0077. Wright-Patterson Air Force Base, OH (2012)
14. Shively, J., Flaherty, S., Miller, C., Fern, L., Neiswander, G.: Delegation control in control of unmanned aerial systems (UAS). In: *Infotech @ Aerospace Conference, AIAA-2012-2458* (2012)
15. Draper, M.H., Miller, C.A., Calhoun, G.L., Ruff, H., Hamell, J., Benton, J., Barry, T.: Multi-unmanned aerial vehicle systems control via flexible levels of interaction: an adaptable operator-automation interface concept demonstration. In: *AIAA Infotech @ Aerospace Conference, AIAA-2013-4803* (2013)
16. Miller, C.A., Draper, M., Hamell, J.D., Calhoun, G., Barry, T., Ruff, H.: Enabling Dynamic Delegation Interactions with Multiple Unmanned Vehicles; Flexibility from Top to Bottom. In: Harris, D. (ed.) *EPCE 2013. LNCS (LNAI)*, vol. 8020, pp. 282–291. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-39354-9_31](https://doi.org/10.1007/978-3-642-39354-9_31)
17. Calhoun, G., Draper, M., Miller, C., Ruff, H., Breeden, C., Hamell, J.: Adaptable automation interface for multi-unmanned aerial systems control: preliminary usability evaluation. *Proc. Hum. Factors Ergon. Soc.* **57**(1), 26–30 (2013)
18. Calhoun, G., Ruff, H., Behymer, K., Frost, M.: Human-autonomy teaming interface design considerations for multi-unmanned vehicle control. *Theor. Issues Ergon. Sci. Special Issue: Human-Autonomy Teaming* (2017, in press)
19. Rowe, A., Spriggs, S., Hooper, D.: Fusion: a framework for human interaction with flexible-adaptive automation across multiple unmanned systems. In: *International Symposium on Aviation Psychology* (2015)
20. Behymer, K., Rothwell, C., Ruff, H., Patzek, M., Calhoun, G., Draper, M., Douglass, S., Kingston, D., Lange, D.: Initial evaluation of the intelligent multi-UxV planner with adaptive collaborative/control technologies (IMPACT). Technical report, AFRL-RH-WP-TR-2017-0011. Wright-Patterson Air Force Base, OH (2017, in press)
21. Bennett, K.B., Flach, J.M.: *Display and Interface Design: Subtle Science*, Exact Art. CRC Press, Boca Raton (2011)
22. Shneiderman, B.: *Designing the User Interface: Strategies for Effective Human-Computer Interactions*, 2nd edn. Addison-Wesley, Reading (1992)
23. Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnicky, A.I.: Pocketsphinx: a free, real-time continuous speech recognition system for hand-held devices. In: *IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings (ICASSP)*, vol. 1, pp. I-185–I-188 (2006)

24. Behymer, K., Rothwell, C., Ruff, H., Patzek, M., Calhoun, G., Draper, M., Douglass, S., Kingston, D., Lange, D.: Second evaluation of the intelligent multi-UxV planner with adaptive collaborative/control technologies (IMPACT). Technical report, AFRL-RH-WP-TR-2017-TBD. Wright-Patterson Air Force Base, OH (2017, in press)
25. Feitshans, G.L., Davis, J.E.: Advanced net-centric combat unmanned air vehicle. Technical report, AFRL-RG-WP-TR-2011-0126, Wright-Patterson Air Force Base, OH (2011)
26. Sears, A., Shneiderman, B.: High precision touchscreens: design strategies and comparisons with a mouse. *Int. J. Man Mach. Stud.* **34**(4), 593–613 (1991)
27. United States Department of Defense: MIL-HDBK-759C: Handbook for Human Engineering Design Guidelines (1995)
28. Shin, G., Zhu, X.: Ergonomic issues associated with the use of touchscreen desktop PC. *Proc. Hum. Factors Ergon. Soc.* **55**(1), 949–953 (2011)
29. Taylor, G.B., Purman, P., Schermerhorn, G., Garcia-Sampedro, R., Hubal, K., Crabtree, A., Rowe, A., Spriggs, S.: Multi-modal interaction for UAS control. In: *SPIE defense + security*, p. 946802. International Society for Optics and Photonics (2015)

Research on User Mental Model Acquisition Based on Multidimensional Data Collaborative Analysis in Product Service System Innovation Process

Jinhua Dou^{1,2} and Jingyan Qin^{1(✉)}

¹ School of Computer and Communication Engineering, School of Mechanical Engineering, University of Science and Technology Beijing, Beijing, People's Republic of China
doujinhua6971@163.com, qinjingyan@gmail.com

² School of Art and Design,
Tianjin University of Technology, Tianjin, People's Republic of China

Abstract. The core of innovation design for product service system is to provide products and services meeting the needs of users. Traditional user research methods have many drawbacks, such as data interference, fuzzy feedback, single-dimensional indicators lack mutual verification, these issues have not been effectively resolved. If the designer can't understand the users' needs in time and objectively, and accurately construct the users' mental model, it will make the design decision blurred and slow, and the design scheme lacks the utility. Taking into account the drawbacks of traditional user research methods, we use the multidimensional data collaborative analysis technology to obtain user, environment, tasks and other context information. The interaction between the user and the product service system is selected, and the context data of the target user is acquired by using the wearable device, the eye movement measurement and behavior analysis system. These methods can assist the designer discover the product service system usability problem, understand the users' needs, build the users' mental model. This paper focuses on the research of user mental model acquisition mechanism in Product Service System Innovation Process. Based on context awareness multidimensional data collaborative analysis, the users' mental model is constructed accurately. To promote the effective matching between the design conceptual model and the user mental model to produce the optimal design, avoid the waste of the design process resources.

Keywords: Product service system · User needs · Mental model · Context awareness · Multidimensional data · Collaborative analysis

1 Introduction

Traditional enterprises are committed to providing tangible products for the people, resulting in a large number of energy wastes in the production, sale, use, recycling process. It also causes a shortage of resources, environmental pollution and other serious problems affecting people's daily lives. With the increasingly serious environmental problems, there is an urgent need to change the traditional product design. The designer

should research sustainable innovation design mechanism which takes into account the needs of users, environmental benefits, social benefits and enterprise development. Product Service System (PSS) is a kind of innovation strategy change under the social sustainable development target [1]. Product service system provides products and services to users. With the rapid development of information technology, tangible products are gradually weakening, product service system presents intelligent and non-material characteristics. The users of Product service system are not only ordinary users, but also disabled, the elderly, children and other primary users. Users, especially the primary users can't be good use of the products and services when they face to dematerialization, information technology and high-tech products service system provided by the enterprise. It will produce a negative user experience when the products and services provided by enterprises can't meet the user needs. Users will eventually abandon the product service system, resulting in waste of design and service process. these presents a huge challenge to designers. Users will generate awareness and unconscious expectations in the process of interaction with product service system, forming a user mental model. Designers change the inherent psychological image into a product service system external form, forming a design concept model [2]. If the designer can't understand user needs and accurate construct the user mental model, it will make design decision-making ineffective. How to obtain the characteristics of users' needs accurately and construct the users' mental model in the process of product service system innovation is the main problem to be solved.

The context of User interaction with the product service system is changing at any time, which impacts the user physiological and psychological, so the designer should consider user needs in a more specific context. Schilit [3] classifies the context as locations, identifications of persons and objects, and changes in these objects. In the process of using product service system, the context data form multi-dimensional space, and there is regularity and correlation between data. The designer can be more intelligent, real-time, accurate access context information Based on context-aware technology in the process of user interaction with the product service system. Designers can objectively understand the context status of the users, find usability issues, mining user needs and build user mental model. The results of this research will provide reference methods for designers to miming users' needs, improve the efficiency and effectiveness of design decision, and to provide more personalized service.

2 Related Works

2.1 User Knowledge and Design Knowledge

User knowledge includes the users' physiological, psychological, cognitive, behavior, social knowledge. Design knowledge includes the designer's experience, thinking, design process and product function, material, shape, i.e. [4, 5]. In the process of design innovation, the user information data is acquired through the scientific research methods, the user mental model is constructed according to the user knowledge and reasoning rules as well as the user knowledge interpretation. Through the users' knowledge acquisition, analysis, reasoning to map the users' mental model to the functional model

and the program model, iterative optimization design program [6, 7]. by this method, the designer can promote the transformation of user knowledge to product knowledge effectively, and finally match the design conceptual model and user mental model effectively to achieve sustainable product service system innovation.

2.2 Mental Model

Psychologist Kenneth Craik [8] first proposed the concept of mental model in 1943. Donald Norman [2] introduced the concept of mental model into the design field, and proposed three models in the design field, namely, system model, design concept model and user mental model.

The design concept model focuses on the designers cognitive understanding things in the creation process. User mental model understand interrelation and Interaction process between the user and product or system. The system model takes into account the overall interaction model and the law of things to run. Mental model help people to discover the laws of things and understanding new things or information, then guide people to deal with various relationships. We can explore the formation process of the mental model in the process of using product service system, that is, how the users form the expectation of consciousness or unconsciousness to the existing product service system according to the knowledge, experience and context. Mental model can also help designer to understand users' perception process and interpretation process when they use product service system, that is, how users perform feature recognition, matching and meaning of activation to design program with prior knowledge.

2.3 The User Research Methods Based on Mental Model Representation

It is important to obtain the user mental model accurately for making a design decision. The user research methods based on mental model representation mainly applied to user domain research of product service system. The methods include Questionnaire method, Interview method, Observation method, Think-aloud method, concept map, Card Sorting. i.e. the designers can understand aimed users' psychological need and construct users' mental model by these methods.

Delugach et al. [9] described a method of direct acquisition of team mental models in the form of conceptual graphs which applied a knowledge capture approach and a supporting graphical tool. S. Angsupanich and S. Matayong [10] developed the prototype of blinds' mobile application. They applied the interview method to obtain the blind user mental model in the process of Product Innovation. E. Kowalczyk and A. Memon [11] applied the GUI testing method to obtain the user mental model. they verified the effectiveness of GUI testing methods by comparing the two mental models of 12 Android apps – one derived from the app's usage and the other from its public description. Lorraine Normore and Vandana Singh [12] described a preliminary study of application to support the user-centered design of future information and communication technologies. A. Pentel [13] applied Think-aloud protocol to connect user emotions and mouse movements. A. Nawaz [14] presents how the choice of card sorting techniques affected the results of the information structure for websites.

These methods are easy to operate. However, there are many subjective elements in questionnaire, interview and observation, which affect the accuracy of the analysis results. Think-aloud method is easy to increase the users' psychological burden. Card Sorting based on the content rather than the task, it is not suitable for the users' real context, which may lead to the analysis results are not accurate enough. Concept mapping needs to refer to specific research objects and contexts, it is not suitable for applying to cognitive information with ambiguous levels or Different styles.

2.4 The User Research Methods Based on Physiological Measurement Data

User research based on physiological measurement data, such as Skin electrical, ECG, EEG, i.e. which help designers to gain a deeper understanding of user experience and internal cognitive processing mechanism in the process of using product service system [15]. Masaki et al. [16] quantified evaluated the software experience using EEG. Mauri et al. [17] revealed that using Facebook could evoke a psychophysiological condition characterized by high valence and high arousal. They used some methods including skin conductance, blood volume pulse, electroencephalogram, electromyography, respiratory activity, and pupil dilation. Ge et al. [18] studied the application and future research prospects of electrophysiological parameters such as skin electrical, ECG and EEG in user experience research. Electrophysiological Technique may apply in the design of products and services such as helping the disabled, medical care and intelligent driving in future. The signal is easily influenced by external environment, physical activity and psychological stimulation when we measure the users' physiological signals, it may interference with accuracy of the data results

2.5 The User Research Methods Based on Eye Movement Experiment

The user research based on eye movement experiment is widely used in the design process of products and services, which is mainly used in the study of consciousness and unconscious demand, usability testing process. Chen et al. [19] developed different 3D animations fostered students to generate better learning performance and sophisticated mental models. They also used eye movements to evaluate users' mental model. Li et al. [20] proposed a method for user requirement acquisition based on eye tracking. Lai Meng-Lung et al. [21] revealed how eye tracking technology was used for learning study. There are also some problems in eye movement experiment, such as user eye movement trajectory can't completely reflect the true idea of the user; it is difficult for researchers to distinguish clearly internal factors simply according with the eye movement data; when the number of users is small, eye movement results may vary greatly. Thus, a single experiment using eye movements to obtain experimental results may have deviations or errors.

2.6 The User Research Methods Based on Virtual Reality Technology

User research based on virtual reality technology can be used to study user needs, function and program evaluation of product service system. And we can understand the users'

experience and behavior more realistically with virtual reality technology. Virtual reality system has been developed and applied in the design, training and ergonomic evaluation applications. Kuliga et al. [22] suggested that virtual reality is a potential practical tool for supporting behavioral validation in psychology, architectural research, and future research. F. Meng and W. Zhang [23] simulated a fire emergency with virtual reality technology and studied the pathfinding behavior in a fire accident. Otebolaku and Andrade [24] explored and evaluated the context-aware smartphone application and recognition classification algorithms. Dong et al. [25] explored the user experience and assessment of context-aware smart home based on virtual reality research methods. There are also some problems in the actual use of virtual reality, such as expensive, poor platform compatibility, which affect the wide application of virtual reality technology in product service innovation. In the process of experimental research, various data acquisition methods have different characteristics.

User psychology is often affected by environmental, task, physiological, psychological, social experience and other context factors. In the design process, the designer should consider a variety of factors in the process of using product service system, explore user behavior, motivation and demand in-depth, accurate obtain user mental model, design the function and program of product service system to avoid invalid design process and useless product service system.

3 Acquisition of User Mental Model Based on Multidimensional Data Collaborative Analysis in Product Service System

The development of Internet of Things technology makes it more convenient to obtain the experimental data. Designers can understand the user context more objectively, accurate mining user mental model based on the multidimensional data collaborative analysis method. At the beginning of the experiment, we first select the target user and build context scenarios. Then we obtain the users' electroencephalogram (EEG), electrocardiogram (ECG), Skin electricity, electromyogram (EMG), respiration, heart rate, pulse and other physiological data in real time through a variety of sensor devices. Through WIFI, GPS technology, light sensors and other physical sensors, we obtain environmental light, noise, temperature, location and other environmental context data. Through the behavior analysis system, we obtain the users' action and behavior data.

The multidimensional data are huge and complex, and the data contains many invalid data, such as noise and outliers, data duplication and missing. We remove the noise and irrelevant data in the data set by data cleaning, and analyze the law of the data itself and the correlation among the data. We simplify the data and find the useful feature to reduced data size and amount of data as much as possible. We mining, calculate and analyze the multidimensional large-scale data. We obtain the user characteristics through clustering, association, artificial neural network and visual analysis. The clustering algorithm gathers the data with same characteristics together. It can help designer find product availability problems and Mining personality or common features. The association rule mining model will acquire the interdependence and correlation of different dimension data, and discover the association

rule, correlation or causal structure through algorithm. We can find the data hidden rules, mining user needs, obtain user mental model by these methods.

4 Acquirement of Elderly User Mental Model in the Process of Using Intelligent TV Product Service System

The aging of the modern society is becoming increasingly serious, the proportion of the elderly population continues to grow. “2015 World Population Prospects” shows that in 2010, the world’s age population of 65-year-old and over 65 years is about 538 million, accounting for 8% of the world’s total population. By 2050, the number of people aged 65 years and over 65 years will reach 1.5 billion, accounting for 16% of the world’s total population. The increase older population has driven the development of a silver economy dedicated to the provision of goods and services for the elderly. With the development of information technology, modern products and service systems have more high-tech features. Product service system is change from the passive service which the user makes a service request to the predictive, perceptual, emotional smart service. Products with good availability service system will give the elderly a pleasant experience and provide human services to elderly. In the process of using the product and service, Elderly users will form the mental model of product service system due to the limitation of experience, knowledge and other factors, as well as physiological and psychological factors. If the designer can’t accurately obtain the elderly users’ mental model, it will make the design concept model and user mental model can’t effectively match, leading to poor availability of product service system. The elderly user is not easy to understand and operate, which affect the efficiency of task completion, trigger the frustration of older users and form negative experience (Fig. 1).



Fig. 1. Mental model measurement of elderly users in real family context

We obtained the older users’ emotion state when they watched smart TV in the real home environment through experimental methods. We used the Changhong 50Q3T smart TV models to set the experimental context. We Selected 30 elders aged from 55 to 75 years old to obtain their mental model for level search and Pinyin search TV program by simultaneous measurement their physiological, psychological and behavior

data. We obtained the following data from a variety of sensor devices: the elders' physiological data including ECG, skin electrical, EMG, breathing, heart rate, pulse and other physiological data. Environmental data was including light data, temperature data and other environmental data. There were also the elders' behavior and activity data in the implementation of the task. We also obtained the elders' facial expression data through a behavior analysis system (Fig. 2).

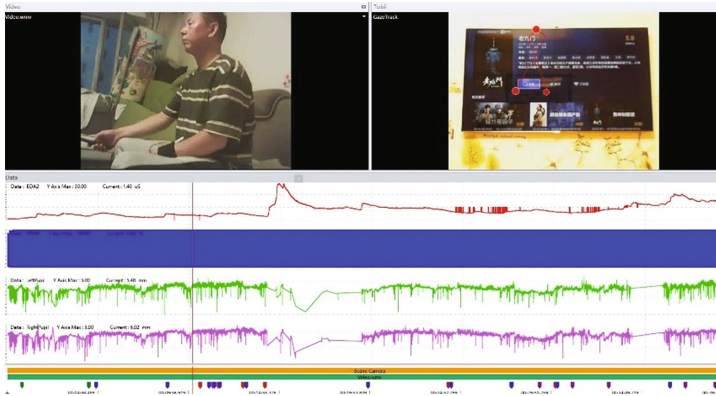


Fig. 2. Multidimensional data acquisition of smart TV for elderly

In this part, the project team analyzed the acquired physiological signals and emotional data, corrected and cleaned up the acquired error data, and converted the multi-source heterogeneous data into computable processed data form. In the algorithm, we mainly used clustering and classification algorithms. Cluster analysis is composed of several patterns. Normally, a pattern is a vector of measurements, or a point in a multidimensional space. Clustering analysis is based on similarity, it has more similarity between patterns in a cluster than patterns that are not in the same cluster. Clustering analysis algorithm is divided into partitioning method, hierarchical method, density-based method, grid-based method. The classification algorithm finds the classification rule by analyzing the training set of the known category, and predicts the classification of the new data. The single classification method mainly includes: decision tree, Bayesian, artificial neural network, K-nearest neighbor, support vector machine and classification based on association rules.

Through the study, we find that most (about 70%) of elderly users have high HF (parasympathetic activity) and low LF (sympathetic activity) in the process of pinyin search. They also have relatively high LF (sympathetic activity) and relatively low HF (parasympathetic activity) in the process of level search. Behavioral data analysis shows activity characteristics for a user to complete a search task with two different search methods. The experimental data show that the average time for a user to complete search task with a level search is significantly less than the average time for a pinyin search. By observing the expression and behavioral characteristics of older users, we find that older users often have frustration emotion in the process of using pinyin search. These multidimensional data indicate that the user experience

of level search is superior to pinyin search. The current TV remote control pinyin input model is different from older users' mental models. Some older people with lower educational levels affect their pinyin input search operations. In addition, older users' poor vision and slowly behavior make pinyin search more difficult. We also make interviews with older users, most older users tend to operate a simple hierarchical search approach. Some older users have suggested that they want simpler or smarter programs to search, such as voice search. But in the process of using voice search, the obstacles arising mainly from the voice isn't standard, it may affect voice search function.

The elderly perception, cognitive and physiological functions showed a downward trend with the increase of age. Elderly vision decline, hearing dropped, behavioral delay, the ability adaptability to various environments gradually weakened, these features highlight the usability problems of product service system. The designer can objectively understand the psychological needs of older users through multidimensional experimental data analysis. They should design product service system easy to use and provide more convenient services to elders.

5 Conclusions

We propose a method for acquiring user mental models based on multidimensional data synchronization analysis. In the process of product service system design, the designer access to the users' mental model objectively and accurately, they should take into account the users' various context factors. By using a variety of sensors, eye tracking and behavior analysis as well as interview, designer can obtain users' multidimensional data information objectively and analysis users' mental model. Through the design knowledge acquisition and reasoning, the mapping relation between the user mental model and the design conceptual model is established, and the iterative optimization design scheme is realized. Finally, the effective match between the user mental model and the design conceptual model is achieved, and the optimal product service system design meeting the user demand is produced.

The multidimensional data collaborative analysis method is applied to the innovative design process of the product service system. It can help designer produced reasonable design schemes to meet the physiological, psychological needs and context characteristics of the users, avoid the waste of resources caused by the unreasonable design schemes. This has certain value for sustainable innovation design theory and product service system design practice.

Acknowledgements. We would like to thank Kingfar International Inc and Qingju Wang, Guoqiang Sun, mengda Yang support our experimental data acquisition. We also thank Ministry of Education Humanities and Social Sciences Research Youth Fund Project and Grant No. 15YJCZH034 that partially supported our research work.

References

1. Manzini, E., Vezzoli, C.: A strategic design approach to develop sustainable product service systems: examples taken from the 'environmentally friendly innovation' Italian prize. *J. Cleaner Prod.* **11**(8), 851–857 (2003)
2. Norman, D.A.: *The Design of Everyday Things*. Basic Books, New York (2002)
3. Schilit, B., Adams, N., Want, R.: Context-aware computing applications. In: *Proceedings of the Workshop on Mobile Computing Systems and Applications*, vol. 16(2), pp. 85–90 (1994)
4. Zhu, S.S.: *Research on product form design technology based on knowledge*. Doctoral dissertation, Zhejiang University (2003)
5. Luo, S.J., Zhu, S.S., Ying, F.T., et al.: Status and progress of research on users' tacit knowledge in product design. *Comput. Integr. Manuf. Syst.* **16**(4), 673–688 (2010)
6. Luo, S.J., Zhu, S.S., Sun, S.Q., et al.: Case study on user knowledge and design knowledge in product form design. *China Mech. Eng.* **15**(8), 709–712 (2004)
7. Liu, Z., Sun, S.Q., Wu, J.F., et al.: Model of product form innovation design knowledge based on users' cognition. *Comput. Integr. Manuf. Syst.* **15**(2), 265–270 (2009)
8. Craik, K.J.W.: *The Nature of Explanation*. Cambridge University Press, Cambridge (1943)
9. Delugach, H.S., Eitzkorn, L.H., Carpenter, S., Utley, D.: A knowledge capture approach for directly acquiring team mental models. *Int. J. Hum.-Comput. Stud.* **96**, 12–21 (2016)
10. Angsupanich, S., Matayong, S.: Applying the mental model for real-time recognition of Thai banknotes: the blinds' mobile application. In: *3rd International Conference on Computer and Information Sciences (ICCOINS)*, Kuala Lumpur, Malaysia, pp. 86–90 (2016). doi:[10.1109/ICCOINS.2016.7783194](https://doi.org/10.1109/ICCOINS.2016.7783194)
11. Kowalczyk, E., Memon, A.: Extending manual GUI testing beyond defects by building mental models of software behavior. In: *30th IEEE/ACM International Conference on Automated Software Engineering Workshop*, Lincoln, NE, pp. 35–41 (2015). doi:[10.1109/ASEW.2015.17](https://doi.org/10.1109/ASEW.2015.17)
12. Normore, L., Singh, V.: Mental models: setting user expectations for ICTs. In: *Proceedings of the 2012 iConference (iConference 2012)*, pp. 550–551. ACM, New York (2012). doi:<http://dx.doi.org/10.1145/2132176.2132284>
13. Pentel, A.: Employing think-aloud protocol to connect user emotions and mouse movements. In: *6th International Conference on Information, Intelligence, Systems and Applications*, Corfu (2015), pp. 1–5 (2015). doi:[10.1109/IISA.2015.7387970](https://doi.org/10.1109/IISA.2015.7387970)
14. Nawaz, A.: A comparison of card-sorting analysis methods. In: *Asia Pacific Conference on Computer-Human Interaction, Apchi 2012* (2012)
15. Dirican, A.C., Göktürk, M.: Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Comput. Sci.* **3**(1), 1361–1367 (2011)
16. Masaki, H., Ohira, M., Uwano, H., Matsumoto, K.: A quantitative evaluation on the software use experience with electroencephalogram. In: Marcus, A. (ed.) *DUXU 2011*. LNCS, vol. 6770, pp. 469–477. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-21708-1_53](https://doi.org/10.1007/978-3-642-21708-1_53)
17. Mauri, M., Cipresso, P., Balgera, A., Villamira, M., Riva, G.: Why is Facebook so successful? psychophysiological measures describe a core flow state while using Facebook. *Cyberpsychol. Behav.* **14**(12), 723–731 (2011)
18. Ge, Y., Chen, Y.N., Liu, Y.F., et al.: Electrophysiological measures applied in user experience studies. *Adv. Psychol. Sci.* **22**(6), 959–967 (2014)

19. Chen, S.C., She, H.C., Hsiao, M.S.: Using eye-tracking to investigate the different 3D representation on students' mental model construction. In: IEEE 15th International Conference on Advanced Learning Technologies, pp. 388–390. Hualien (2015). doi:[10.1109/ICALT.2015.150](https://doi.org/10.1109/ICALT.2015.150)
20. Li, Z., Gou, B.C., Chu, J.J., et al.: Way of getting user requirements based on eye tracking technology. *Comput. Eng. Appl.* **51**(9), 233–237 (2015)
21. Lai, M.L., Tsai, M.J., Yang, F.Y., Hsu, C.Y., Liu, T.C., Lee, W.Y., et al.: A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educ. Res. Rev.* **10**(4), 90–115 (2013)
22. Kuliga, S.F., Thrash, T., Dalton, R.C., Hölscher, C.: Virtual reality as an empirical research tool — exploring user experience in a real building and a corresponding virtual model. *Comput. Environ. Urban Syst.* **54**, 363–375 (2015)
23. Meng, F., Zhang, W.: Way-finding during a fire emergency: an experimental study in a virtual environment. *Ergonomics* **57**(6), 816–827 (2014)
24. Otebolaku, A.M., Andrade, M.T.: User context recognition using smartphone sensors and classification models. *J. Netw. Comput. Appl.* **66**, 33–51 (2016)
25. Dong, W.S., Kim, H., Kim, J.S., Lee, J.Y.: Hybrid reality-based user experience and evaluation of a context-aware smart home. *Comput. Ind.* **76**, 11–23 (2016)

Are 100 ms Fast Enough? Characterizing Latency Perception Thresholds in Mouse-Based Interaction

Valentin Forch^{1(✉)}, Thomas Franke², Nadine Rauh¹,
and Josef F. Krems¹

¹ Department of Psychology, Cognitive and Engineering Psychology,
Chemnitz University of Technology, Chemnitz, Germany
valentin.forch@psychologie.tu-chemnitz.de

² Institute for Multimedia and Interactive Systems, Engineering Psychology
and Cognitive Ergonomics, University of Lübeck, Lübeck, Germany

Abstract. The claim that 100 ms system latency is fast enough for an optimal interaction with highly interactive computer systems has been challenged by several studies demonstrating that users are able to perceive latencies well below the 100 ms mark. Although a high amount of daily computer interactions is still characterized by mouse-based interaction, to date only few studies about latency perception thresholds have employed a corresponding interaction paradigm. Therefore, we determined latency perception thresholds in a mouse-based computer interaction task. We also tested whether user characteristics, such as experience with latency in computer interaction and interaction styles, might be related to inter-individual differences in latency perception thresholds, as results of previous studies indicate that there is considerable inter-individual variance in latency perception thresholds. Our results show that latency perception thresholds for a simple mouse-based computer interaction lie in the range of 60 ms and that inter-individual differences in latency perception can be related to user characteristics.

Keywords: Latency · System response time · Human-computer interaction · Mouse-based interaction · Latency perception

1 Introduction

Optimizing system latency (the time delay between user input and the output response of a computer system [13, 16]) is a common challenge in any interactive computer system. System latency can degrade user experience [2, 18] and lead to a less efficient interaction with a given system [10, 11, 13, 17]. Consequently, specifying design goals in terms of acceptable latencies has been a fundamental research topic in human-computer interaction for several decades [4, 13, 14]. Ideally, an interactive computer system should mimic physical systems as closely as possible to allow for a fluid and natural interaction. Hence, an optimal computer system should have no subjectively perceptible time delay between system input (e.g., hand movements) on the one side and respective system output (e.g., cursor movement) on the other side.

Determining this latency perception threshold (also termed just noticeable difference, JND [16]) means to find the system latency where users cannot distinguish between a system with and without additional latency anymore.

In the past, common latency design guidelines typically proposed the design goal of 100 ms system latency for an optimal interaction with highly interactive computer systems [5, 14]. However, recent research [6, 16] has shown that users can perceive much lower latencies (around 10 ms) when interacting with a touchscreen device (i.e., direct-touch interaction). Studies applying indirect input paradigms suggest that users can detect latencies as low as 50 ms when using a touchpad or a stylus [3, 6]. For mouse-based interaction, first research indicates that latencies below 100 ms can impair performance [10, 17]. While a high amount of daily computer interactions is still characterized by mouse-based interaction, there are only few studies about latency perception thresholds employing a corresponding interaction paradigm. Therefore, it appears relevant to validate these recent results with this prevalent interaction type.

Furthermore, previous studies have provided evidence for the existence of considerable inter-individual differences in latency perception thresholds. Annett and colleagues [3] found latency perception thresholds in the range of 30–80 ms and 60–105 ms for stylus-based tasks (drawing and writing). An even greater range of 20–100 ms was reported by Jota, Ng, Dietz, and Wigdor [11] for a tapping task where participants had to evaluate the time delay between tapping a touch display and the appearance of a rectangle. Explaining this variability of latency perception thresholds found across users in different tasks should be informative for latency perception studies as well as latency design guidelines, as the average latency perception threshold of a population is of relatively little value when the population variance is high. One way of explaining this variance might be its relation to inter-individual differences in user characteristics. Two classes of variables of particular interest in this case might be the previous experience with latency in computer interaction [20], and inter-individual differences in interaction styles [19].

The objective of the present research was twofold – first, we aimed to determine the magnitude of latency perception thresholds in a mouse-based human-computer interaction task. Second, we intended to explore factors which may lead to inter-individual differences in these identified latency perception thresholds.

2 Background

2.1 Perception of System Latency

Echoing Annett and colleagues [3]: “From a psychological and interaction perspective, it is imperative to understand the processes governing latency perception before recommendations for future systems are made” (p. 173). One way of shedding light on the underlying processes of latency perception is the comparison of latency perception thresholds across different tasks. In experimental tasks where participants solely relied on the temporal offset between an input and an output (e.g., tapping on a touch screen and waiting for a change of the display), latency perception thresholds have been found to be considerably higher compared to tasks where latencies lead to changes on more

salient dimensions, such as the spatial offset between finger position and cursor position in direct-touch interaction [16]. Abstracting from this, perceiving relatively low system latencies seems to require the comparison of the input state and the output state of a computer system along dimensions influenced by system latency that are readily perceivable (e.g., spatial information in contrast to temporal information).

The spatial offset between the input device and the cursor in direct-touch pointing or dragging tasks which is caused by system latency, can be directly processed in the visual system [15, 16]. Compared to this, the perception of latency in mouse-based pointing or dragging tasks should require a more complex comparison. This is because information about the input and the output are distributed across different perceptual modalities (somatosensory system and visual system respectively) and have to be transformed before being compared. More distributed representations of the input state and the output state should therefore hinder their comparison, effectively leading to higher latency perception thresholds. Indeed, Annett and colleagues [3] found that making visual information about the input state (but not the output state) of a system unavailable increased latency perception thresholds in an inking task.

While the available information about the input and the output state of a system fundamentally shape how latency is perceived, there are other factors which should also determine if a user will perceive latency when performing a given task. Following Annett and colleagues' [3] latency perception model, two main factors influencing latency perception are contextual demands (e.g., task requirements, environmental factors) and the observer of the system. Observer characteristics relevant for latency perception may include experience of how latency manifests in computer interactions, as well as practice with tasks highly sensitive to latency. If the observer of the system also provides the input of the system, characteristics of the interaction, such as movement speed of the input device controlled by the user, should be of importance for latency perception, as they may lead to a higher saliency of the system latency. Hence, differences in movement speed and experience with tasks where latency detection is relevant may be used to explain inter-individual variance of latency perception thresholds.

2.2 Movement Speed

Higher movement speeds in dragging or pointing tasks cause a larger spatial offset between the input device and the cursor when latency is present, because the input device travels a farther distance before the cursor position is updated and this change is displayed [15, 16]. In direct-touch dragging tasks the spatial offset between the finger and the center of the dragged object is still perceivable at latencies as low as 10 ms when moving at a moderate pace [15]. This means that even at very low latencies an increase in movement speed should increase the spatial offset between the hand and the cursor and therefore make it more likely that users will perceive latency. This relationship between movement speed and latency perception should be the same for mouse-based interaction, albeit being less pronounced, because of the more complex comparison of mouse position (i.e., input state) and cursor position (i.e., output state). While direct-touch dragging allows to continuously compare finger position and cursor

position, evaluating the simultaneousness between mouse position and cursor position may only be possible when initiating or changing the direction of a movement (i.e., when the cursor has once started to follow a movement vector of the hand it is difficult to estimate the displacement distance to the position of the mouse).

2.3 Experience with Latency in Computer Interaction

Apart from comparing input state and output state of a system, users might also draw from their experience about how latency manifests in the computer interaction to compare the actual output state of the system with an expected output state (i.e., if there was no perceptible system latency). A common activity where users might gather experience about the impact of system latency and practice latency detection are highly dynamic computer games, such as action games, racing games, or first person shooter games. Because these games require very fast and precise reactions depending on the output state of the system, they are especially susceptible to system latency, so even small system latencies can impair performance. This is also a reason why these games are a relevant research subject for studying the influence of latency on user performance [10].

2.4 Research Questions and Hypotheses

The purpose of the present study was to answer two questions – first, we aimed to determine the magnitude of latency perception thresholds in a mouse-based human-computer interaction task.

Q1: Of which magnitude are the latency perception thresholds in a mouse-based human-computer interaction task?

Our second research question concerning the inter-individual differences of latency perception thresholds was:

Q2: Are there user characteristics related to inter-individual differences in latency perception thresholds?

In respect of the second research question we hypothesized that higher movement speeds should lead to a higher discrepancy between expected and actual mouse cursor position and therefore to a higher perceptibility of latency, resulting in a lower latency perception threshold of individuals with higher average mouse movement speeds.

H1: Higher mouse movement speeds are related to lower latency perception thresholds.

We also expected individuals with a higher exposure to highly dynamic computer games, such as action games, racing games, or first person shooter games, to exhibit lower latency perception thresholds, as they might have a higher sensitivity towards a mismatch of actual and predicted mouse cursor position.

H2: More experience with highly dynamic computer games is related to lower latency perception thresholds.

3 Method

3.1 Participants

Twenty students (10 female, age 19–36 years, $M = 23.45$, $SD = 3.32$) which were recruited via the local psychology student mailing list took part in the experiment. All participants had normal or corrected-to-normal vision and normally used their right hand for handling computer mice. Participants signed an informed consent sheet at the beginning of the experiment and received partial course credit for participation.

3.2 Experimental Setup

Procedure. Participants were asked to complete a mouse-based dragging task on a computer screen. The task was to move a grey square representing the mouse cursor from the left side of the display to the right and back again without pausing. The left and right target areas were each indicated by two short dashes at the bottom and the top of the display (see Fig. 1). Both target areas covered 20% of the display. Moreover, participants were allowed to touch the left and right edges of the screen. Therefore, no precise dragging was necessary to perform the task which means that the task difficulty (i.e. index of difficulty according to Fitts' Law [8]) and only a relatively low level of coordination was required (i.e., low workload for action control).



Fig. 1. Dragging task. Participants were instructed to move the grey square representing the mouse cursor from the left to the right side of the screen and back again without pausing. Arrows were not present in the experiment.

To determine participants' latency perception thresholds we used an adaptive threshold estimation approach (ZEST – for more details regarding this method see section “Scales and Measures”) in a two-alternative forced-choice discrimination task (2AFC task). Hence, each trial consisted of two subtrials showing (a) the *reference system* with baseline latency and (b) the *probe system* with additional latency ranging between 1 and 300 ms. The systems were presented in a randomized order. After finishing both subtrials, participants were asked to indicate in which subtrial the system reacted instantaneously, that is without additional latency. The latency perception threshold was defined as the additional latency of the probe system where participants were able to correctly distinguish between the reference system and the probe system 75% of the time. This is a commonly accepted perception threshold value for 2AFC tasks lying between the maximum hit rate of 100% and the baseline hit rate of 50% (the guessing probability when performing at chance level) [16, 21].

To ensure that participants were familiar with the task and understood how added latency becomes apparent, they completed a short training with a fixed and relatively high probe latency of 200 ms. The training ended when five trials in a row were answered correctly. It took participants five to eight trials to finish the training. The experimental trials were divided into two separate ZEST runs providing two perception threshold estimates for each participant. Each ZEST run consisted of 30 trials, except for the first three participants who took part in the study where the number of ZEST trials was still set to 20 trials. While this is a small inconsistency in our methodology that should be considered in interpreting our results, the effect on the results should be small and should, at the most, only lead to a somewhat lower precision of our estimates (i.e., instead of biasing results in a certain direction). However, it also has to be noted that there is no linear relationship between the number of ZEST trials and reliability of determining the latency threshold (i.e., the effect of increasing the number of ZEST trials from 20 to 30 is relatively small).

Hardware and Software. The hardware setup consisted of a computer with a 3.3 GHz processor (Intel(R)Core™ i7-5820k), a mouse with a polling rate of 1000 Hz and a sensor resolution set to 800 dpi (Logitech G303), and a 24-inch-monitor with a refresh rate of 144 Hz (Acer XF240H). The experimental task was implemented in C++ using the open-source libraries of SFML (www.sfml-dev.org) for 2D visual displays. System latency was manipulated by delaying the update of the mouse cursor position for a fixed amount of time throughout the probe subtrials.

System latency. To measure the latency of the computer system we recorded the mouse and the monitor with a 1000 Hz high speed camera (Sony RX100IV) while running the program of the experimental task and initializing a sharp movement of the mouse with a solid object [10]. The latency of the system was then calculated from the number of frames elapsed between the start of the mouse movement and a first update of the mouse cursor position. We measured the baseline system latency and the system latency with additional latency to assess the accuracy of the latency manipulation ten times each. The mean baseline latency of the system amounted to 8 ms ($SD = 1.7$ ms). Adding 200 ms latency through the software resulted in a mean system latency of 207 ms ($SD = 1.5$ ms) differing only slightly from the expected system latency of 208 ms when considering the baseline of 8 ms.

3.3 Scales and Measures

Estimation of Latency Perception Thresholds. The Bayesian threshold estimation approach ZEST [12] allows to perform exact threshold estimations within as little as 20 trials [1]. Starting from a hypothetical prior distribution of perception thresholds, this procedure estimates the perception threshold (and therefore the optimal latency of the probe to test for this threshold) based on the participant's performance in each prior trial. This means that a correct (incorrect) answer at a given latency level leads to a decrease (increase) of the participant's estimated perception threshold which in turn is used as the latency level of the probe for the next trial. The update of the prior distribution is done via multiplication with a likelihood function which gives the probability of a correct (or incorrect) answer depending on the participant's current estimated perception threshold and the latency of the probe [1].

We used the Weibull function as our likelihood function, which is also dependent on the fixed parameters for the guessing rate (γ), the lapsing rate (λ), the slope factor (β) and the parameter ε . Because we used a 2AFC task, γ was set to .50 (the hit rate when performing at the level of chance), λ was set to .03, β was set to 3.5 as suggested in [22] and ε was set to -0.03 (following Eq. 4 from [1]). The prior distribution was a normal distribution on a log scale centered on 2 with a standard deviation of 0.7 – resulting in a prior distribution mean of 100 ms (10 to the power of 2) and the inclusion of thresholds from 20 ms (≈ 10 to the power of 1.3) up to 500 ms (≈ 10 to the power of 2.7) within ± 1 SD of the mean. The upper bound for the probe latency for each trial was set to 300 ms, because previous studies indicate that users should readily perceive system latencies of this magnitude [3, 6] and higher system latencies are not representative for normal computer systems [10].

Movement speed. We measured participants' movement speed in two ways. The first measure was the average movement speed across all probe subtrials, the second measure was the average covered distance after the first 100 ms of movement for each probe subtrial focusing on the first acceleration of each subtrial. Both measures were computed from a log file of the cursor position on a millisecond time scale. We included the second measure to focus on the first displacement of the mouse cursor, where we assumed additional latency to be best visible, as well as to control for the possibility that some participants might move slower in certain sections, which would in turn systematically decrease their average movement speed.

Experience with highly dynamic computer games. To assess the amount of the participants' practice with highly dynamic computer games we asked for the time they typically spent playing highly dynamic computer games, such as action games, racing games, or first person shooter games within a week. We asked this question with regard to the past six months (six month experience), as well as the period in the participants' lives when they were most actively playing these games (lifetime experience).

4 Results

4.1 Magnitude of Latency Perception Thresholds

With regard to our first research question (Q1) we computed measures of central tendency and dispersion of participants' latency perception thresholds. Because the retest reliability of the two latency perception threshold estimates of each participant was high ($r = .87$), their mean was used for subsequent analyses. The latency perception thresholds' range was 34–137 ms with a mean of 65 ms (*Median* = 54 ms) and a standard deviation of 30 ms. The latency perception threshold distribution was right-skewed and a Shapiro-Wilk test indicated that the latency perception thresholds were non-normally distributed ($W(20) = 0.88, p = .021$).

4.2 Inter-individual Differences Related to Latency Perception Thresholds

To address our second research question (Q2) we aimed to quantify the strength of the associations between latency perception thresholds and average movement speed, average covered distance after 100 ms, as well as six month, and lifetime experience with highly dynamic computer games. We decided to not analyze the six month experience with highly dynamic computer games, because 60% of the participants did not play any of these games in the past six months causing a substantial restriction of variance for this variable. The percentage of participants without any lifetime experience with highly dynamic computer games was only 30%. The average time spent playing highly dynamic computer games within a week in the period of the participants' lives where they were most actively playing these games was 10.25 h (*Median* = 4, *SD* = 12.73). Because of the non-normality of the latency perception threshold distribution we used non-parametric Spearman correlations.

To test our first hypothesis (H1) we computed the correlation between latency perception thresholds and average movement speeds, as well as average covered distance after 100 ms. Both higher average movement speeds and higher average covered distance after 100 ms were associated with lower latency perception thresholds ($r_s = -.23, r_s = -.35$ respectively), but the correlations did not reach statistical significance ($p > .05$).

As proposed in our second hypothesis (H2), a higher lifetime experience with highly dynamic computer games significantly correlated with participants' latency perception threshold values ($r_s = -.42, p = .031$). To further illustrate the effect of the lifetime experience with highly dynamic computer games we split the sample into two groups along the median of this variable. The mean latency perception thresholds of the groups with a high and low lifetime experience with highly dynamic computer games were $M_{\text{High Exp.}} = 52$ ms ($SD_{\text{High Exp.}} = 6$ ms), and $M_{\text{Low Exp.}} = 78$ ms ($SD_{\text{Low Exp.}} = 11$ ms) respectively.

5 Discussion

Our results indicate that on average users are able to perceive latencies around 60 ms (combining the median latency perception threshold and our baseline system latency) in mouse-based interaction tasks. We also found considerable inter-individual differences in latency perception thresholds in our sample with some participants reaching or even surpassing the 100 ms mark. These results come close to values reported in recent studies where participants performed other forms of indirect input tasks [3, 6, 11]. Furthermore, we found that observer characteristics can play a role in latency perception, as the inter-individual differences in latency perception thresholds were related to the experience with highly dynamic computer games, such as action games, racing games, or first person shooter games. The average latency perception threshold of the group with little prior experience with highly dynamic computer games was about 25 ms higher than the thresholds of the high-experience group.

There was also a tendency for participants with higher movement speeds to have lower latency perception thresholds, but the effect was too small to reach statistical significance. While others have argued that an increased spatial offset between input device and cursor caused by higher movement speeds does not increase the saliency of latency [3], we would not rule out this possibility given our results and with our small sample size in mind, which restricted the power to detect such a relatively small effect. Albeit the difference between our two measures of average movement speed is small, we found a greater effect of movement speed on latency perception thresholds for the measure of covered distance after 100 ms, which only included information about the first left-to-right movement in our experiment. We would argue that there might be types of movements or specific movement phases (e.g., movement onset, direction changes) where movement speed could be relevant for latency perception, but studies with a higher power would be needed to detect this effect.

5.1 Practical and Theoretical Implications

The most important practical implication of the present study might be that latency perception thresholds can be associated with user characteristics. This means that when measuring latency perception thresholds or designing a computer system based on the results of these measurements, one should also consider the user characteristics of the sample the measurements were conducted with, the user characteristics of the target audience, as well as the fit of these two groups with respect to their relevant user characteristics.

While our results give support to the notion that user characteristics play an important role in latency perception in principle, our experiment does not allow to pinpoint the mechanisms that caused this effect. We theorized that the experience with highly dynamic computer games should make it more likely for users to perceive lower latencies, as these games implicitly provide the user with some information about how latencies manifest in computer interaction. Other mechanisms possibly mediating the relationship between experience with highly dynamic computer games and latency perception thresholds might be the participants' motivation to perform well in a

computer interaction task, aspects of the interaction style apart from pure movement speed, or cognitive abilities, such as spatial perception.

Another confounding variable could be the latency of the computer systems participants interact with on a day-to-day basis. As highly dynamic computer games have higher hardware requirements compared to normal office software or web browsers, users who play (or played) these games also might own a faster computer system to match these hardware requirements. It is possible that participants of our study might have referred to their usual latency experience as a baseline when doing the 2AFC task. It also has to be noted that our study only provided correlational data which does not necessarily imply a strong causality between experience with highly dynamic computer games and latency perception thresholds.

5.2 Future Research

Firmly establishing the connection of latency perception and user characteristics cannot be achieved with a single study, especially when taking into account the relatively small sample size of the present one. Therefore, subsequent studies are needed to test this relationship and possible mediators, such as mentioned above. Apart from probing the relationship of experience with highly dynamic computer games, subsequent studies might also focus on gathering more data about interaction styles, such as movement speed or the role of strategies for latency compensation, which can also involve the adaptation of movement speeds [7, 9]. Yet another avenue for future research might be the influence of user's normal latency experience when interacting with their own computer systems. As was pointed out before [16, 20], user's expectations about how much latency a system should exhibit are highly influenced by the technical standard they gather their experiences with, which is continuously improving. These expectations could also be related to user's latency perception or effect it indirectly through altered interaction styles.

6 Conclusion

The results of our study provide some new insights for the understanding of latency perception. First, we found that latency perception thresholds for a simple mouse-based task lie well below 100 ms, similar to the results of studies about latency perception thresholds for other indirect input tasks [3, 6, 11]. However this does not necessarily mean that users will perceive latencies below 100 ms or find them disturbing [3]. In the light of the available information about latency perception thresholds the recommendation of 100 ms system latency [5, 14] should nevertheless be scrutinized in further studies [4]. Second, we demonstrated that inter-individual differences in latency perception are related to user characteristics, in this case the experience with highly dynamic computer games. We also argued that interaction characteristics such as movement speed might play a role in latency perception, although further studies are needed to investigate this effect. Together these findings open up many possibilities for future research about latency perception, particularly the role of user characteristics and interaction styles.

Acknowledgements. This research was funded by the German Federal Ministry of Education and Research (03ZZ0504H) in the context of the project fast-realttime. Statements in this paper reflect the authors' views and do not necessarily reflect those of the funding body or of the project partners.

References

1. Alcalá-Quintana, R., García-Pérez, M.A.: The role of parametric assumptions in adaptive Bayesian estimation. *Psychol. Methods* **9**, 250–271 (2004). doi:[10.1037/1082-989X.9.2.250](https://doi.org/10.1037/1082-989X.9.2.250)
2. Anderson, G., Doherty, R., Ganapathy, S.: User perception of touch screen latency. In: Marcus, A. (ed.) DUXU 2011. LNCS, vol. 6769, pp. 195–202. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-21675-6_23](https://doi.org/10.1007/978-3-642-21675-6_23)
3. Annett, M., Ng, A., Dietz, P.H., Bischof, W.F., Gupta, A.: How low should we go? Understanding the perception of latency while inking. In: *Proceedings of Graphics Interface 2014*, pp. 167–174. Canadian Information Processing Society (2014)
4. Attig, C., Rauh, N., Franke, T., Krems, J.F.: System latency guidelines then and now – is zero latency really considered necessary? In: *Paper Presented at HCI International 2017* (2017)
5. Card, S.K., Robertson, G.G., Mackinlay, J.D.: The information visualizer, an information workspace. In: *CHI 1991 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 181–186. ACM (1991). doi:[10.1145/108844.108874](https://doi.org/10.1145/108844.108874)
6. Deber, J., Jota, R., Forlines, C., Wigdor, D.: How much faster is fast enough? User perception of latency & latency improvements in direct and indirect touch. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1827–1836. ACM (2015). doi:[10.1145/2702123.2702300](https://doi.org/10.1145/2702123.2702300)
7. de la Malla, C., Lopez-Moliner, J., Brenner, E.: Dealing with delays does not transfer across sensorimotor tasks. *J. Vis.* **14**, 8 (2014). doi:[10.1167/14.12.8](https://doi.org/10.1167/14.12.8)
8. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. *J. Exp. Psychol.* **47**, 381–391 (1954). doi:[10.1037/h0055392](https://doi.org/10.1037/h0055392)
9. Honda, T., Hirashima, M., Nozaki, D.: Adaptation to visual feedback delay influences visuomotor learning. *PLoS ONE* **7**, e37900 (2012). doi:[10.1371/journal.pone.0037900](https://doi.org/10.1371/journal.pone.0037900)
10. Ivkovic, Z., Stavness, I., Gutwin, C., Sutcliffe, S.: Quantifying and mitigating the negative effects of local latencies on aiming in 3d shooter games. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 135–144. ACM (2015). doi:[10.1145/2702123.2702432](https://doi.org/10.1145/2702123.2702432)
11. Jota, R., Ng, A., Dietz, P.H., Wigdor, D.: How fast is fast enough? A study of the effects of latency in direct-touch pointing tasks. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2291–2300. ACM (2013). doi:[10.1145/2470654.2481317](https://doi.org/10.1145/2470654.2481317)
12. King-Smith, P.E., Grigsby, S.S., Vingrys, A.J., Benes, S.C., Supowit, A.: Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. *Vis. Res.* **34**, 885–912 (1994)
13. Mackenzie, I.S., Ware, C.: Lag as a determinant of human performance in interactive systems. In: *Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems*. pp. 488–493. ACM (1993). doi:[10.1145/169059.169431](https://doi.org/10.1145/169059.169431)
14. Miller, R.B.: Response time in man-computer conversational transactions. In: *Proceedings of the December 9–11, 1968, Fall Joint Computer Conference, Part I*, pp. 267–277. ACM (1968). doi:[10.1145/1476589.1476628](https://doi.org/10.1145/1476589.1476628)

15. Ng, A., Dietz, P.H.: The effects of latency and motion blur on touch screen user experience. *J. Soc. Inform. Display* **22**, 449–456 (2014). doi:[10.1002/jsid.243](https://doi.org/10.1002/jsid.243)
16. Ng, A., Lepinski, J., Wigdor, D., Sanders, S., Dietz, P.H.: Designing for low-latency direct-touch input. In: *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, pp. 453–464. ACM (2012). doi:[10.1145/2380116.2380174](https://doi.org/10.1145/2380116.2380174)
17. Pavlovych, A., Gutwin, C.: Assessing target acquisition and tracking performance for complex moving targets in the presence of latency and jitter. In: *Proceedings of Graphics Interface 2012*, pp. 109–116. Canadian Information Processing Society (2012)
18. Potter, J.J., Singhose, W.E.: Effects of input shaping on manual control of flexible and time-delayed systems. *Hum. Factors* **56**, 1284–1295 (2014). doi:[10.1177/0018720814528004](https://doi.org/10.1177/0018720814528004)
19. Seow, S.C.: *Designing and Engineering Time: the Psychology of Time Perception in Software*. Pearson Education, Boston (2008)
20. Shneiderman, B., Plaisant, C.: *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, Boston (1987)
21. Ulrich, R., Vorberg, D.: Estimating the difference limen in 2AFC tasks: pitfalls and improved estimators. *Atten. Percept. Psychophys.* **71**, 1219–1227 (2009). doi:[10.3758/APP.71.6.1219](https://doi.org/10.3758/APP.71.6.1219)
22. Watson, A.B., Pelli, D.G.: QUEST: a Bayesian adaptive psychometric method. *Atten. Percept. Psychophys.* **33**, 113–120 (1983). doi:[10.3758/BF03202828](https://doi.org/10.3758/BF03202828)

Design and Evaluation of an Assistive Window for Soft Keyboards of Tablet PCs that Reduces Visual Attention Shifts

Bomyeong Kim¹, Kyungdoh Kim²(✉), Jinho Ahn²,
and Robert W. Proctor³

¹ Graduate Program in Cognitive Science, Yonsei University, Seoul, Korea
ghaud31@nate.com

² Department of Industrial Engineering, Hongik University, Seoul, Korea
kyungdoh.kim@hongik.ac.kr, h2oinjan@gmail.com

³ Department of Psychological Sciences, Purdue University, 703 Third St.,
West Lafayette, IN 47907, USA
proctor@psych.purdue.edu

Abstract. In general, soft keyboards are used for tablet PCs much as they are for smartphones. However, since the screen size is larger than for smartphones, user gaze is dispersed. In this paper, we propose a new keyboard for tablet PCs, in which an assistive window is introduced to decrease visual shifts, and evaluate the usability and usefulness of the keyboard. Results of an experiment are reported confirming that the gaze dispersion showed a reduction of 90% compared to an existing keyboard, although the typing error rates and typing time did not greatly improve. These results indicate a way to reduce visual dispersions on larger screen devices.

Keywords: Tablet PC · Soft keyboard · Visual attention shifts · Typing error rates

1 Introduction

Recently, the use of tablet personal computers (PCs) has steadily risen, and applications for them have been continuously developed [2, 29]. However, the tablet PC has been used only as a device for consuming information rather than producing it because its input devices are insufficient compared to those of desktop and laptop PCs [15, 26]. To make better use of a tablet PC, it is necessary to improve the existing soft keyboards.

Most recent soft keyboard designs for mobile devices have had the goal of reducing typing errors that are a consequence of faulty touch [4, 18, 20] or decreasing muscle fatigue of the hands [6, 19, 22]. Research to date has been mostly aimed at improving the soft keyboard that operates on a small screen like that of a mobile phone [3, 16]. But tablet PCs have larger screens than mobile phones do, as illustrated in Fig. 1. Consequently, when typing on soft keyboards, visual shifts occur more often and the distance between the typed message and the keyboard is larger. This greater distance between the keyboard



Fig. 1. Distances between the cursor and keyboards: (a) 5.5 inch Mobile Phone, (b) 8.3 inch Tablet PC.

and positions on the screen should result in more eye fatigue than on a smartphone, as well as other negative effects [24, 27]. Eye movements interfere with a person’s visual working memory, which could disrupt a user’s task performance [11, 21]. Consequently, the user would spend more time visually searching for a target key and moving the responding hand to the target position [27]. This interference should have a negative effect on keyboard usage, such as typing speed and typing errors [27].

In the present study we propose a new type of soft keyboard for tablet PCs that can reduce the fatigue and typing error rates associated specifically with visual shifts between the keyboard and typed message. We also report a study evaluating the keyboard’s usability.

2 Literature Review

A number of researchers have evaluated the typing performance and user satisfaction of existing keyboards [5, 7, 17]. Kim and Park [17] analyzed and evaluated the usability of seven smartphone soft keyboards designed for Korean. Similarly, Cuaresma and MacKenzie [7] compared the four keyboards using QWERTY layout. As a result, they showed that the “Octopus” (implemented by K3A, <http://ok.k3a.me/>), which provided the prediction of the words on top of the next letter, was the fastest. Chaparro et al. [5] compared soft keyboards with physical keyboards. They found that participants typed faster with the physical keyboard, but also committed more typing errors.

In addition, there has been research that compared different keypad sizes [25, 29]. With stylus input, MacKenzie and Zhang [25] reported that a smaller keypad increased errors but did not reduce speed compared to a larger keypad. In contrast, with multiple finger input, Sears and Zha [29] showed that keypad sizes did not affect data entry speeds, and making the keypad smaller did not increase error rates or negatively impact preference ratings.

Previous work found that users who were experienced with desktop systems were better with the QWERTY layout [5, 7, 23, 25]. Therefore, they recommended using the QWERTY layout soft keyboards on Tablet PCs.

The lack of tactile responses on soft keyboards is well-documented and makes it difficult for users to get key-click confirmations during text entry with soft keyboards [12, 13]. To solve this problem, Han and Kim [13] showed that the typing performance was improved by adding tactile feedback to the existing visual display on a commercial tablet and phone.

Additionally, there were attempts to design a new keyboard layout by rearranging split-keys [9, 23, 28]. MacKenzie and Zhang [23] designed a keyboard layout applying a model which predicted the upper-bound text entry rates for soft keyboards, and considered the shortest path [30]. Schoenleben and Oulasvirta [28] introduced a keyboard design that allows ten-finger touch typing by utilizing a touch sensor on the back side of a device. The new layout was unfamiliar to users. After training, the new keyboard typing was faster than a QWERTY keyboard. Also, Go and Tsurumi [9] used existing QWERTY layout, and added the text entry option of the pie menu as the selection method to overcome the limited size of soft keyboard.

Many studies have examined typing patterns on soft keyboards [1, 8, 10, 14, 16, 18]. Hirche et al. [14] proposed a novel approach to text input to resolve the inherent difficulties with text input on mobile phone. They used a very limited set of buttons that, by using word prediction and hints, would only require minimal finger movements. Similarly, Findlater and O. Wobbrock [8] introduced two novel personalized keyboard interfaces, both of which adapted their key-press classification models. In addition, since the cause of many typing errors on soft keyboards is that their buttons are too small, improving input efficiency by predicting words typed by users and changing the button size dynamically have been studied [1, 8, 16, 18].

However, several studies not only evaluated the typing performance, but also measured the physical discomfort when typing with soft keyboards [6, 22]. Except for the study of Paek [27], however, most studies focused on decreasing muscle fatigue of

the hands or arms. According to the study of Paek [27], gaze movement between the on-screen keyboard and the text entry area affects input performance because it gives inefficiency and extraneous workload to the user. To reduce gaze movement, methods such as Glass Pad (placing the keyboard on the text area as a semi-transparent state and moving along the cursor when entering text) or String Cursor (displaying strings on the keyboard entered by user) were developed [27]. However, because those methods are based on use of a pen to enter characters, the results may not be useful when entering characters by hand.

Most prior studies of soft keyboards either evaluated the typing performance of soft keyboards, proposed a new type of keyboard, or examined a method to decrease muscle fatigue of the hands or arms. These studies demonstrate a lack of research for reducing the dispersion gaze when using soft keyboards on tablet PCs. Therefore, we aim to focus on decreasing the dispersion gaze, then, we offer a new soft keyboard design intended to reduce dispersion gaze and evaluate whether it in fact does so. The remainder of this paper is organized as follows. In Sect. 3, we propose a new soft keyboard design and explain the specific experimental method. In Sect. 4, we report a usability study that evaluates how much improved the proposed new soft keyboard is in performance and preference compared to a traditional keyboard. Lastly, in Sect. 5, we conclude and discuss future research.

3 Method

3.1 New Soft Keyboard Design

As mentioned, the cause of visual attention shifts when using the soft keyboard is that the window in which the text is displayed and the keyboard for entering text are not in the same area. Therefore, we designed a new soft keyboard to shorten the distances between the text window and the soft keyboard for entering text. The new soft keyboard is based on a QWERTY layout, which is the most commonly used, with an assistive window added onto the soft keyboards, different from existing keyboards (Fig. 2). This assistive window immediately displays the texts entered through the soft keyboard. In other words, the entered texts are shown at the top of the soft keyboard so that users can recognize which texts are being typed without looking at the text window. This closer proximity should reduce user's eye dispersions since users do not have to look at the text field.

Additionally, we designed the assistive window so that color of the background is black and color of text is yellow, in order to make the typed strings conspicuous [27]. Furthermore, through showing all currently typed text lines when entering strings, it is possible to identify typographical errors easily from typing without the user's eyes having to move to the actual text window. And the assistive window expands to show all of the typed text until user presses the enter button.

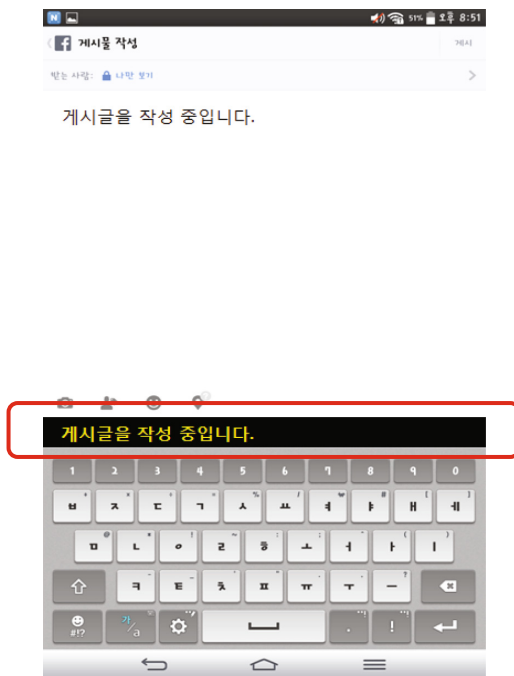


Fig. 2. The new soft keyboard with an assistive window (The window is highlighted with a red-line surround). (Color figure online)

3.2 Experimental Design

There is a necessity to receive assessment and feedback about how much eye dispersions are actually reduced by the new soft keyboard. To perform such an assessment, we conducted an experiment with five male and five female Korean university students whose mean age was 23.8 years. The device that we used is Gpad, an Android 8.3 inch tablet PC (Table 1). Seven participants were majoring in Industrial Engineering, one in Mechanical Engineering, one in Computer Engineering, and one in English language. 90% of the participants had used smartphones more than three years. Seven of the participants had used the QWERTY keypad on their smartphones, and seven had experience of using tablet PCs. And we used a QWERTY keyboard in Korean for experiment.

The experiment was performed over two days (sessions), and one session consisted of two tests. In each test, the same set of 20 sentences was presented to participants with different keyboard settings. Participants were instructed to memorize each sentence and type it as quickly and correctly as possible, without reading it again. All presented sentences were selected randomly from among 300 catchy short sentences such as proverbs and brief lines in movies. Two sets of sentences (set A and set B) were provided. The sentences of Set A and B were selected randomly from among 300 sentences for each session, so the sentences of Set A and B were different for each session. The order of presenting the sentences for the two keyboard tests was different.

Table 1. Device descriptions

Model	LG G pad 8.3
Display	8.3 inch IPS LCD 1920 × 1200 pixel Touch screen
Size (WxHxD)	216.8 × 126.5 × 8.3 mm
Camera	5 MP/1.3 MP
Weight	338 g

One set (set A) was used in the first session, and the other set (set B) in the second session. The proposed new soft keyboard was used on one test, and the existing soft keyboard was used on the other test in the same session. The keyboard order was counterbalanced across the two sessions for each participant to control for order effects. Furthermore, the typing speed (Dependent Variable (DV1, unit: Characters per minute, CPM) and typing error rate (DV2, unit: %) were measured for each test. The error rates were calculated by dividing the sum of the number of backspace keystrokes and of touching the text field to modify the typing errors into total amounts of characters in a sentence. Also, the experimental tests were videotaped to measure the number of eye movements (DV3, unit: number). An experimenter watched the tapes and recorded an eye movement each time gaze movement between the on-screen keyboard and the text entry area occurred, similar to the study of Paek [27].

After each test, rating scores were gathered. Participants evaluated the keyboard on which they had just been tested with 3 questions about its usefulness, 4 questions about usability, and 1 question pertaining to overall satisfaction. The questions were seven-point Likert items ranging from 1 being mostly disagree to 7 being mostly agree. To test the internal consistency of the ratings, one paired question (Questions 4 and 8 about eye strain) was included in the questionnaire. At the end of each session, a total of 8 questions were also assessed by participants in order to get user feedback about the new soft keyboard. Lastly, at the end of all tests, a “scenario analysis” was performed by showing 7 scenarios that were similar to real situations and asking each participant which keyboard s/he would want to use in each scenario.

The procedure was as follows. Participants’ informed consent, including for video recording the tests, was given at the beginning. Then, in the first step, the participants were asked to provide their demographic information and were informed about the overall experimental procedure. After that, the first session was performed. When each test was completed, the evaluation ratings were gathered. At the end of the session, participants were asked to provide feedback about the new soft keyboard. The second session was conducted similarly to the first session after one or a few days. Finally, the “scenario questions” were administered at the end of the second session; they were designed to determine whether the proposed new soft keyboard was regarded as preferable in several contexts.

4 Results

4.1 Typing Speed, Errors, and Visual Shifts

An analysis of variance (ANOVA) for the typing speeds showed no difference between the two keyboards, $F(1, 9) = 1.95, p = 0.196$. The error rates also showed no difference, $F(1, 9) < 1.0$. Most important, as shown in Fig. 3, the number of visual shifts was markedly diminished when using the new soft keyboard, $F(1, 9) = 54.25, p < .0001$.

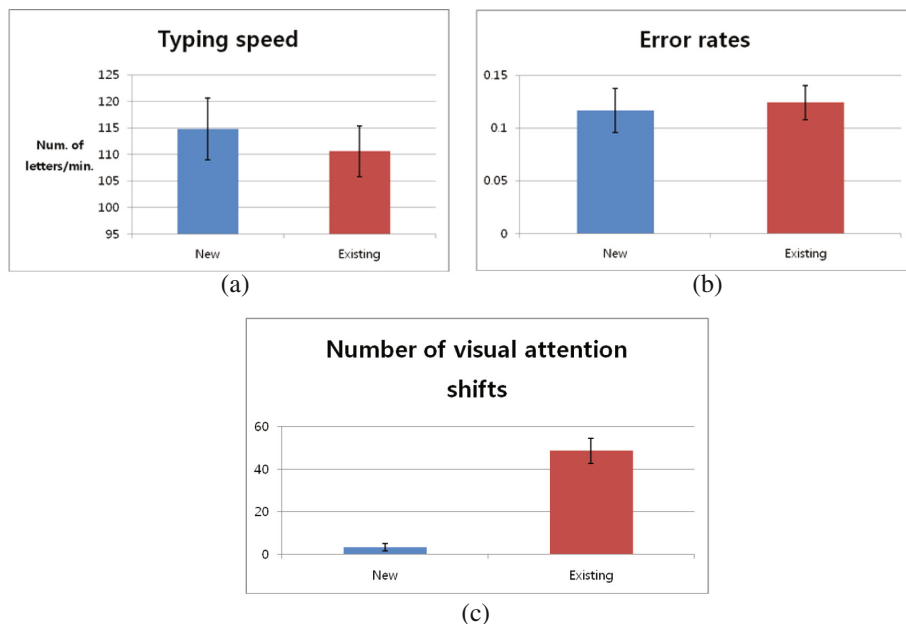


Fig. 3. Graphs of (a) typing speed, (b) error rate, and (c) number of visual shifts (Error bars are standard error of the mean).

4.2 Usability, Usefulness and Satisfaction

ANOVAs conducted for each question showed that there were significant differences between the new soft keyboard and the existing soft keyboard on all questions except for question 3. Besides, the level of reliability on the paired question was high, Cronbach's Alpha equal to 0.939. The results of each question are shown in Table 2.

According to these results, participants agreed more strongly that they could type accurately with the new soft keyboard than with the existing soft keyboard (Question 1). Although error rates were not improved (see Sect. 3.1), the ratings for low probability of mistyping were higher with the new soft keyboard (Question 2). However, participants did not indicate any advantage for the new soft keyboard over the existing keyboard for typing fast (Question 3).

Table 2. Summary of rating scores

Aspects	Questions	New keyboard	Existing keyboard	F value (df = 9)	p-value
Usefulness	1. I could type accurately with this soft keyboard without errors.	5.1	4.0	14.83	0.0039
	2. The probability of mistyping was low with this soft keyboard.	5.0	3.5	7.80	0.0210
	3. I could type fast with this soft keyboard.	4.9	4.8	0.17	0.6911
Usability	4. My eyes got easily tired when I use this soft keyboard.	3.0*	4.0	16.36	0.0029
	5. It was easy to notice typing errors with this soft keyboard.	6.0	3.2	56.00	<.0001
	6. I noticed typing errors quickly with this soft keyboard	6.0	3.2	63.43	<.0001
	7. There were many visual attention shifts when I use this soft keyboard.	2.4*	5.2	40.09	0.0001
	8. My eyes were tired when I use this soft keyboard	3.3*	4.1	9.26	0.0140
Satisfaction	9. I am satisfied with this soft keyboard.	5.9	4.0	76.53	<.0001

Note: Bold font indicates which keyboard was rated higher (or lower with *negative questions) for significant differences

Usability of the new soft keyboard showed better results than the traditional soft keyboard. The participants indicated that their eyes were more fatigued with the existing keyboard than the new keyboard (Question 4). When using the new soft keyboard, they agreed more strongly that they could find errors easily (Question 5) and quickly (Question 6). Moreover, the ratings showed much stronger agreement with the statement that there were many visual shifts for the existing soft keyboard than for the new soft keyboard (Question 7), which was borne out by the performance results in Sect. 3.1.

Also, there were meaningful differences between the new soft keyboard and the existing soft keyboard for the ratings of usefulness, $F(1, 9) = 11.18, p = 0.009$, and usability, $F(1, 9) = 16.21, p = 0.003$, averaged across the relevant questions, and satisfaction, $F(1, 9) = 76.53, p < .0001$. When using new soft keyboard, the scores of all aspects were higher than using the existing one, as shown in Fig. 4. As a result, we could find that the proposed soft keyboard achieved the goal such as typing accuracy, speed, and less discomforts.

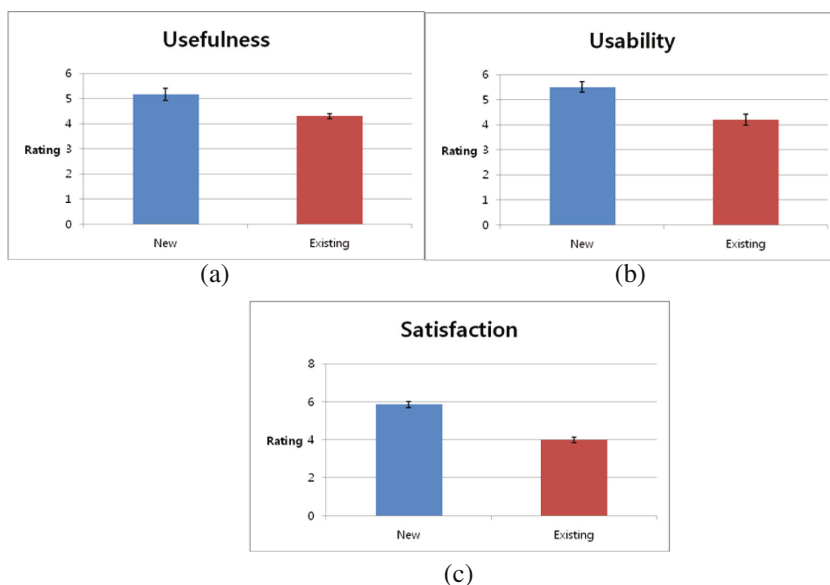


Fig. 4. Rating scores of (a) usefulness, (b) usability, and (c) satisfaction (Error bars are standard error of the mean).

4.3 User Feedback

The average score of each question was over 5.0 on the 7 point scale, except for question 4 (Table 3).

The new soft keyboard was judged as helpful to the participants (Question 1), and the texts in assistive window were rated as conspicuous and showed improved readability (Question 2). Furthermore, most of participants responded that they could find

Table 3. User feedback scores of new soft keyboard

Questions	Mean
1. Using assistive window was helpful to decrease eyestrain when typing the sentences.	5.4
2. The words on the assistive window stood out so it was easy to read the typing words.	6.0
3. I found the typing errors easily through the assistive window.	5.8
4. Thanks to the assistive window, the typing errors were decreased.	4.1
5. The assistive window was located on the appropriate place.	5.7
6. When typing, the assistive window irritated my eyes.	2.9*
7. I put eyes more often on the assistive window than the input window.	5.7
8. I could find the typing errors faster with the assistive window.	6.3

* Negative question

typing errors on the assistive window more easily (Question 3) and faster (Question 8). However, they did not indicate that it decreased typing errors (Question 4).

The typing was somewhat faster even if typing error rates were not improved. This might be attributable to the typing errors being detected quickly thanks to the assistive window. The participants also replied that the new soft keyboard was not unpleasant to the eyes (Question 6) and was located on the appropriate place (Question 5), so they looked at the assistive window more often than text field (Question 7). As a result, the total average score over all of eight questions was 5.3, indicating that the new soft keyboard was perceived as effective.

4.4 Scenario Preferences

Seven kinds of scenarios were presented to the participants, and we asked them to choose which keyboard type they wanted to use in each scenario. The preference results are shown in Table 4.

As a result, most of the participants responded that the new soft keyboard is better in cases of ‘writing long documents such as a report’, ‘writing e-mails with 5 to 10 lines to professor’ and ‘typing a web address’ which got more than 5 votes. According to them, the new soft keyboard can guarantee typing accuracy.

However, they responded that the existing soft keyboard is better in the case of ‘writing a message for messenger’ because most of people generally understand friends’ tiny faults and text fields are already on its soft keyboard in some messengers.

Table 4. Preference results by each scenario

Scenarios	Number of choices		Main comments by individuals
	New keyboard	Existing keyboard	
1. Writing long documents such as a report	7	3	The new keyboard was more useful for long documents.
2. Writing e-mails with 5 to 10 lines	9	1	The new keyboard was more useful for long documents.
3. Writing for SNS (Facebook, Blog etc.)	5	5	The new keyboard was not preferred since typing errors can be ignored.
4. Writing a message for messenger	2	8	The new keyboard was not preferred since typing errors can be ignored.
5. Making a simple note	5	5	The new keyboard was not preferred since typing errors can be ignored.
6. Typing a web address	10	0	The new keyboard was easier and more accurate.
7. Typing an ID and password	5	5	The new keyboard was not preferred since password might be exposed on the assistive window.

For ‘writing for social networking sites (SNSs) like a blog or Facebook’; some participants said that the new soft keyboard is better because people should post carefully on SNS, whereas others answered that the new soft keyboard is not necessary because SNS is a site for communicating with friends similarly to messengers.

Meanwhile, in the case of ‘making a simple note’, the number of choices for the new and existing keyboards did not differ. One group said that the new soft keyboard is not necessary because typing errors are not important in notes, and the other group said that the new keyboard is preferred since the typing accuracy of their notes is important for later use. In the last question, ‘typing in an ID and password’, there was an opinion that the new soft keyboard would have a benefit because this case requires accuracy, but there also was the negative point that passwords can be exposed by the assistive window.

5 Conclusion

We designed a new soft keyboard to shorten the distance between the text window and the soft keyboard for entering text. An assistive window was added just above the soft keyboard. The entered texts were shown in the assistive window so that users could recognize which texts were being typed without looking at the text window. This assistive window reduced user’s eye dispersions.

The following conclusions can be drawn on the basis of the results of the experiment. First, the newly proposed soft keyboard did not significantly increase typing speed. There were not any differences of typing error rates, but typing was somewhat faster with the new soft keyboard. This is because the typing errors can be detected more quickly and easily with the assistive window, although there were no differences on typing error rates. The ability to detect and correct errors would make participants think that the typing error rates were lower with the new soft keyboard.

Second, the new soft keyboard greatly reduced the number of visual attention shifts from fixating on the keyboard to the screen. Unlike the existing soft keyboard for which visual shifts occurred on average of 50 times, the proposed new soft keyboard required them an average of 5 times, a reduction of 90% against the existing one. The reason is that users could check typing strings through the assistive window just above the keyboard so they did not need to move their gaze to the text field. Also, many users responded that ‘the new soft keyboard helps me to reduce eye fatigues’ thanks to the reduced visual shifts.

Third, the users answered that they need the proposed new soft keyboard when accuracy is required because it has the strength of allowing them to detect errors easily through the assistive window. Overall satisfaction with the new soft keyboard was also higher than with the existing soft keyboard. Participants answered that when they used the new soft keyboard, they ‘could type more accurately without error’ and ‘typing error rate was less during typing’. So, we could know the user satisfaction in aspects of usefulness was increased through the three questions. Also, the first two questions indicated that users felt that typing error rates were decreased since they detected typing errors more quickly. Additionally, the new soft keyboard’s scores about usability questions over eye fatigue, visual shifts, and detecting error were higher.

The user feedback results indicate that the new soft keyboard was viewed positively. The participants responded that they gazed at the assistive window more often than the text field. Consequently, the window was helpful to reduce eye fatigue, comfortable to see the text, and easy to find typing errors. We could know that the new soft keyboard proposed in this paper was designed well toward reducing user's eye fatigue by reducing the number of visual attention shifts. Also, as intended, users generally looked at the assistive window of the new soft keyboard rather than the actual text field. In addition, the new soft keyboard drove improvement in aspects of usefulness and usability. Nevertheless, the typing error rates were not reduced; the sentences used in the study may have been too short to show this benefit, so an additional experiment with long sentences needs to be implemented in future work. And we did not include word recommendation. Word recommendation is just above the soft keyboard, but to consider only the effect of the assistive window we did not include word recommendation. As noted above, word recommendation has been applied in the basic soft keyboards of smartphone manufacturers such as Apple, Samsung, and LG. Therefore, if word recommendation is added in our new soft keyboard, we could expect improvement in usability such as typing speed and error rate.

In summary, a soft keyboard with an assistive window, which can minimize visual attention shifts while typing on tablet PCs, was proposed to support producing text content in tablet PC environments. Enhancing the ability to produce content in tablet PCs may require not only an improved soft keyboard but also a variety of supported input devices and applications according to content types. However, the present study is meaningful because the soft keyboard is inevitable on tablet PCs, and the tablet PCs with the assistive window can be used in a wider application area since it minimizes shifts of visual fixation and reduces eye fatigue. Moreover, the new keyboard can also be used for research material that can effectively reduce visual shifts on a variety of devices in addition to tablet PCs.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (Grant No. 2015R1C1A1A01053529).

References

1. Bertram, R.L., Champion, D.F., Hartman, M.E.T.: Computer programmed soft keyboard system, method and apparatus having user input displacement. U.S. Patent, no. 5,818,451. U.S. Patent and Trademark Office, Washington, DC (1998)
2. Bora, K.: PC sales to recover in 2015 while tablet market continues to mature worldwide: Gartner. *International Business Times* (2014). <http://www.ibtimes.com/pc-sales-recover-2015-while-tablet-market-continues-mature-worldwide-gartner-1621596>. Retrieved 8 July
3. Bouteruche, F., Deconde, G., Anquetil, E., Jamet, E.: Design and evaluation of handwriting input interfaces for small-size mobile devices. In: *Proceedings of the 1st Workshop on Improving and Assessing Pen-Based Input Techniques*, pp. 49–56 (2005)

4. Byun, J.: Improvement of computer keyboard design through behavioral analysis - focused on the characteristics of Korean Alphabet and its use. *Korean Soc. Des. Sci.* **15**(3), 241–248 (2002)
5. Chaparro, B.S., Phan, M.H., Jardina, J.R.: Usability and performance of tablet keyboards microsoft surface vs. Apple iPad. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **57**(1), 1328–1332 (2013). Sage Publications
6. Choi, B., Park, S., Jung, K.: Analysis of perceived discomfort and EMG for touch locations of a soft keyboard. In: Stephanidis, C. (ed.) *HCI 2013. CCIS*, vol. 373, pp. 518–522. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-39473-7_103](https://doi.org/10.1007/978-3-642-39473-7_103)
7. Cuaresma, J., MacKenzie, I.S.: A study of variations of Qwerty soft keyboards for mobile phones. In: *Proceedings of the International Conference on Multimedia and Human-Computer Interaction-MHCI*, pp. 126.1–126.8 (2013)
8. Findlater, L., Wobbrock, J.: Personalized input: improving ten-finger touchscreen typing through automatic adaptation. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 815–824. ACM, May, 2012
9. Go, K., Tsurumi, L.: Arranging touch screen software keyboard split-keys based on contact surface. In: *CHI 2010, Extended Abstracts on Human Factors in Computing Systems*, pp. 3805–3810. ACM, April 2010
10. Gunawardana, A., Paek, T., Meek, C.: Usability guided key-target resizing for soft keyboards. In: *Proceedings of the 15th International Conference on Intelligent User Interfaces*, pp. 111–118. ACM, February 2010
11. Gunter, R.W., Bodner, G.E.: How eye movements affect unpleasant memories: support for a working-memory account. *Behav. Res. Ther.* **46**(8), 913–931 (2008)
12. Han, B.-K., Kim, K., Yatani, K., Tan, Hong Z.: Text entry performance evaluation of haptic soft QWERTY keyboard on a tablet device. In: Auvray, M., Duriez, C. (eds.) *EUROHAPTICS 2014. LNCS*, vol. 8618, pp. 325–332. Springer, Heidelberg (2014). doi:[10.1007/978-3-662-44193-0_41](https://doi.org/10.1007/978-3-662-44193-0_41)
13. Han, B., Kim, K.: Typing performance evaluation with multimodal soft keyboard completely integrated in commercial mobile devices. *J. Multimodal User Interfaces* **9**, 1–9 (2015)
14. Hirche, J., Bomark, P., Bauer, M., Solyga, P.: Adaptive interface for text input on large-scale interactive surfaces. In: *3rd IEEE International Workshop on Horizontal Interactive Human Computer Systems, 2008, TABLETOP 2008*, pp. 153–156. IEEE, October 2008
15. Hoggan, E., Brewster, S.A., Johnston, J.: Investigating the effectiveness of tactile feedback for mobile touchscreens. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1573–1582. ACM, April 2008
16. Hong, T., Goh, B., Kim, K.: Adaptive soft keyboard via key-sequence prediction: a development case study on android platform. In: *Proceedings of the Conference of the Korean Institute of Information Scientists and Engineers*, pp. 1767–1769 (2014)
17. Kim, H., Park, B.: Usability evaluation of android-based smart phone soft keyboard. In: *Proceedings of Conference of the Korean Institute of Industrial Engineers*, pp. 1–6 (2010)
18. Kim, S., Kim, N., Byun, W., Choi, J., Kim, T.: Soft keyboard application for reducing the mis-typing ratio in the smartphones. *Korean Inst. Inf. Scientists Eng.* **38**(2), 89–92 (2011)
19. Ko, K., Kim, H.S., Woo, J.H.: The study of muscle fatigue and risks of musculoskeletal system disorders from text inputting on a smartphone. *Ergon. Soc. Korea* **32**(3), 273–278 (2013)
20. Kwon, O.: Implementation of input suggestion system using neighbor miss touch correction method on touch screen smartphones. *Korean Inst. Inf. Scientists Eng.* **39**(1D), 67–69 (2012)
21. Lawrence, B.M., Myerson, J., Abrams, R.A.: Interference with spatial working memory: an eye movement is more than a shift of attention. *Psychon. Bull. Rev.* **11**(3), 488–494 (2004)

22. Lee, K.: Comparison of upper extremity muscle activity with transverse plane angle changes during vertical keyboard typing. *Korean Res. Soc. Phys. Ther.* **16**(2), 67–76 (2009)
23. MacKenzie, I.S., Zhang, S.X.: The design and evaluation of a high-performance soft keyboard. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 25–31. ACM, May 1999
24. MacKenzie, I.S., Zhang, S.X., Soukoreff, R.W.: Text entry using soft keyboards. *Behav. Inf. Technol.* **18**(4), 235–244 (1999)
25. MacKenzie, I.S., Zhang, S.X.: An empirical investigation of the novice experience with soft keyboards. *Behav. Inf. Technol.* **20**(6), 411–418 (2001)
26. Müller, H., Gove, J., Webb, J.: Understanding tablet use: a multi-method exploration. In: Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services, pp. 1–10. ACM, September 2012
27. Paek, J.: Design of Soft Keyboards for Reducing Visual Attention Shifts. School: Information and Communications University (2008)
28. Schoenleben, O., Oulasvirta, A.: Sandwich keyboard: fast ten-finger typing on a mobile device with adaptive touch sensing on the back side. In: Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 175–178. ACM, August 2013
29. Sears, A., Zha, Y.: Data entry for mobile devices using soft keyboards: understanding the effects of keyboard size and user tasks. *Int. J. Hum.-Comput. Interact.* **16**(2), 163–184 (2003)
30. Soukoreff, W.R., Scott MacKenzie, I.: Theoretical upper and lower bounds on typing speed using a stylus and a soft keyboard. *Behav. Inf. Technol.* **14**(6), 370–379 (1995)

Integrated Information Visualization and Usability of User Interfaces for Safety-Critical Contexts

Sonja Th. Kwee-Meier^(✉), Marion Wiessmann,
and Alexander Mertens

Institute of Industrial Engineering and Ergonomics,
RWTH Aachen University, 52062 Aachen, Germany
s.meier@iaw.rwth-aachen.de

Abstract. Safety-critical systems are often designed from a technical point of view. This technocentric approach causes usability problems and, hence, hinders the protection or restoration of safe conditions, such as in emergency response, or even implies safety risks. A systematic literature review was conducted to identify key aspects for a more human-centered design for higher usability and safer system use. General emergency management and response literature as well as study results from various other fields, such as police work, firefighting and nuclear power plants were analyzed. Communalities and differences were contrasted and important overarching design recommendations for safety-critical user interfaces were deduced.

Keywords: Usability · User interface · Human-centered · Safety · Security · Safety-critical · Emergency · Risk

1 Introduction

The importance of usability is widely acknowledged. However, not all systems and interfaces are designed correspondingly. Especially in the field of safety-critical systems, the far-reaching importance of usability and the often low market potential in this field are facing each other. In the work routine, lack of usability can lead to discontent and negative attitudes towards what is supposed to be assistive software. Regarding safety management, the risk of operational mistakes with major consequences increases in high-stress situations as a lack of usability can no longer be outbalanced by enhanced attentional resources. Safety-critical contexts necessitate that information can be easily retrieved by the operators. High usability standards for human-machine interfaces facilitate efficient information access, for example to gain the overview of a safety-critical situation and, in general, to reduce the mental strain in highly demanding emergency situation.

1.1 Psychological Stress of Emergency Personnel

In potentially hazardous situations emergency forces experience increased psychological stress. Negative emotional stress like anxiety, for instance anxiety because of

far-reaching operation errors, reduce the processing efficiency of information (Processing Efficiency Theory, PET [13]). Processing efficiency is to be distinguished from performance efficiency, because the latter can be maintained on a relatively high level up to a certain extent, enabled by compensation mechanisms when there are enough resources like time. However, also the performance efficiency decreases when this limit is exceeded [13]. Based on the PET Eysenck et al. [14] demonstrated the further impact of anxiety on the control of attention in the Attentional Control Theory, ACT. Figure 1 shows the enhanced positive effect of lowered task difficulty in high-stress situations according to the Yerkes-Dodson law [48] on the relationship between arousal and performance [47].

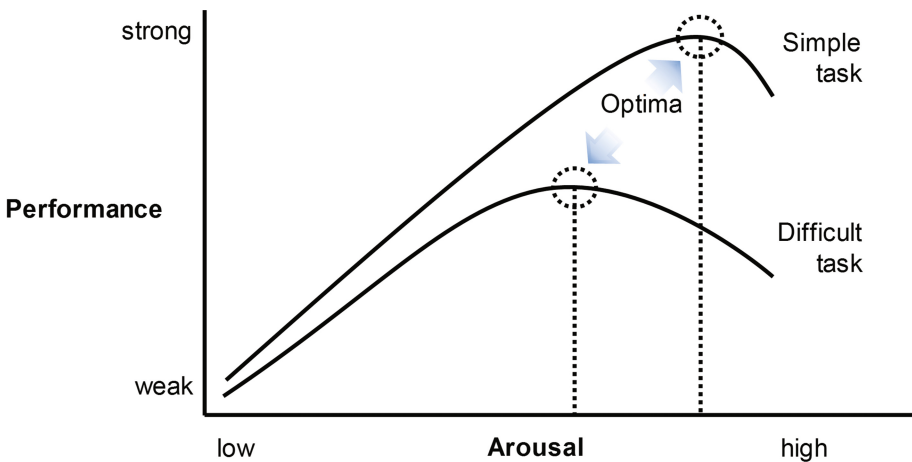


Fig. 1. Inverted U-shaped curves according to the Yerkes-Dodson law [48] regarding the relationship between arousal and performance in dependence on task difficulty (based on Wickens et al. [47]).

Helpful reviews on stress can be found, for instance, in the highly exhaustive overview by Staal [42], depicting the relations between stress, cognition and human performance, the long established textbook on engineering psychology by Wickens et al. [47], and the neuroscientific overview on decision-making in high-stress situations by Starcke [43].

1.2 Motivation and Method for the Present Work

This paper is motivated by but not limited to a key topic in safety emerging over the past years, that is the threat of terrorism. Every item such as a suitcase that appears to be unsupervised and abandoned nowadays becomes easily suspected to contain improvised explosive devices (IEDs). Instructions to build IEDs can be easily accessed by everyone via the Internet. In most cases, the content of suspicious luggage turns out harmless but in rare cases IEDs are indeed contained posing a high risk to the

emergency services and also to every other person in close distance. Therefore, the emergency services are obliged to examine every such item with the utmost care and caution. The safest solution for the emergency services is a remote-controlled system, for instance in form of a robot, which allows them to maintain a safe distance to the suspicious item. At the same time, this distance creates new challenges. The emergency services need to efficiently retrieve information about the item, with the information being fully reliable and offering them a basis to decide on how to proceed in dependence on their conclusion whether or not the content is dangerous.

The literature in the specific field of usability of user interfaces for detecting IEDs is limited. Therefore, a systematic literature review was performed for human-computer interaction in safety-critical contexts in general, covering various fields. Six positive categories and one negative exclusion superordinate category with single terms were formed. The six positive categories were usability (usability, design, usability problems, operational, mistake, error, human factors engineering, information processing, decision support), interface (interface, human-machine, human-computer, person-machine, user, technology, computer), emergency (emergency, emergency case, emergency situation, safe/safety, secure/security, safety-critical system, disaster, crisis, terror/terrorism, hazard, risk, danger), fields (police, defuse bombs, bomb disposal, mitigation measure, improvised explosive devices, fire service/fire fighters, surgery/surgeons, nuclear, power plant), personnel (emergency personnel, security personnel, action force) and stress (panic, fear, mental stress, psychological stress). The focus was on the collaboration between human and technology. Therefore, the exclusion category was defined as collaboration comprising the terms team, communication, collaborative, teamwork. The principles and approaches for user interface design in general and specifically for safety-critical situations were contrasted, communalities merged and recommendations derived for user interfaces for safety-critical contexts.

2 Principles of Usability for Safety-Critical Contexts: Literature Review

In addition to the general usability literature and fundamental guidelines (see Sect. 2.1), 14 theoretical and empirical studies regarding usability of safety-critical systems were identified in the cross-categorical literature search. The implications for safety-critical systems in general are elaborated for the five theoretical and literature based studies in Sect. 2.2, for the three studies with regard to user interface design in firefighting in Sect. 2.3, for the four studies with regard to nuclear power in Sect. 2.4, and for a cross-sectional study integrating avalanche, firefighting, ambulance and police work and for a study with regard to disaster relief operations in Sect. 2.5.

2.1 Fundamental Guidelines for Usability in General

ISO 9241-110:2006 [22] defines seven general principles to be taken into consideration for human-machine dialogue design (see Table 1). These principles can partly be tracked back to the literature guidelines such as the ten usability heuristics by

Nielsen [33] and the “Eight Golden Rules” for interface design by Shneiderman [40]. For instance, the recommendation of conformity with user expectations in the DIN EN ISO 9241-110 [22] can be followed back to the term consistency in Nielson [33] and Shneiderman [40] referring to the same goal. Yet, there are differences acknowledgeable over time, such as from error prevention with emphasis on help and documentation [33] over error prevention with easy reversal of actions [40] to error tolerance [22]. Completeness is not yielded at in this section as the particular focus of this lies on specifically safety-critical systems.

Table 1. Fundamental guidelines for usability.

	Type	Recommendations
DIN EN ISO 9241-110 [22]	Principles for dialogue design	<ul style="list-style-type: none"> • Suitability for the task • Self-descriptiveness • Controllability • Conformity with user expectations • Error tolerance • Suitability for individualization • Suitability for learning
Nielsen (1994) [33]	“Ten Usability Heuristics”	<ul style="list-style-type: none"> • Visibility of system status • Match between system and real world • User control and freedom • Consistency and standards • Error prevention • Recognition rather than recall • Flexibility and efficiency of use • Aesthetic and minimalistic design • Help users recognize, diagnose, and recover from errors • Help and documentation
Shneiderman (1998) [40]	“Eight Golden Rules” for interface design	<ul style="list-style-type: none"> • Consistency • Universal usability • Informative feedback • Dialogue design yielding at closure • Prevention of errors • Easy reversal of actions • Support of internal locus of control • Reduction of short-term memory load

2.2 Theoretical and Literature Based Studies with Regard to Usability for High-Stress Situations

While there is a large amount of literature on general interaction design, such as Sharp [39], usability, such as Welker [46] and user experience, such as Sauro [38], formative studies considering high-stress implications for human-computer interaction (HCI) are sparse. Szalma and Hancock [44] provide an overview of stress theory and its historical development, identifying the positive evaluation of the user interface characteristics as

Table 2. Theoretical and literature based studies.

	Area	Main recommendations
Szalma and Hancock (2008) [44]	Task loading and stress in HCI	<ul style="list-style-type: none"> • Operator training until skill automaticity • Promotion of fast information extraction • Minimum requirements of memory workload
Hancock and Szalma (2003) [19]	Jobs of high task workload and situational stress	<ul style="list-style-type: none"> • Minimize information dispersal over multiple sources • Link new information to currently processed data • Simple display design with simple graphics • Centralized information in an overall view and single figure
Turoff et al. (2004) [45]	Emergency response	<ul style="list-style-type: none"> • Specific training • System adaptability • Self-explanatory design • Hierarchical structured system directory • Easy access to data and information • Immediate data updates • Data as one unit of information
Rahman et al. (2012) [37]	Sociotechnical systems	<ul style="list-style-type: none"> • Emergency controls should not be reduced to touch screens, push buttons, or comparable controls • Emergency controls should be compatible to the operator's perceptual, cognitive and motoric abilities in a stress situation • Emergency controls should adapt to skills

the key factor for coping with the demands following appraisal theories based on Lazarus and Folkman [29] and Lazarus [27, 28]. Lazarus and Folkman [29] framed in the appraisal theory the view of stress as outcome of the personal evaluation of environmental demands, i.e. the person-environment transactions. Szalma and Hancock [44] outline general principles for stress mitigation in human-computer interaction present requirements for further research. One suggested possibility to mitigate stress in an emergency situation is the training of the operators until automaticity (see Table 2). The approach focuses the skill level. Section 1.1 discussed the potentially impaired information processing efficiency under negative emotions according to the PET [13]. Szalma and Hancock [44] take up on this aspect and refer to the reduced ability of solving complex problems and analytical thinking under stress, pointing to the relevance to keep the effort of the working memory to a minimum and enable the operator to extract information as fast as possible.

In a prior study, Hancock and Szalma [19] had developed a study on jobs which have to be performed in the presence of high task workload and situational stress. Based on a literature review, they had derived a number of general principles or guidelines for display design for these conditions. They recommend to minimize information dispersal over multiple sources when working under stressful conditions and to link new information to data in the moment it is being processed, because operators under stress might

neglect new information. Displays that ask for data transformation should be avoided in favor of data presentations with simple graphics [19]. An efficient way to alleviate stress is to require as little information processing as possible. The simplicity of the display design was named as important possibility to facilitate the perception in a stressful situation. In line, simple graphics were demanded for faster comprehension as they reduce the effort for the working memory and support direct instructions [19], resulting at best in centralized information in an overall view and single figure. The temporal dimension should be further captured into an integrated display to counteract the temporal distortion in stress situations and simplify the extraction of information, the temporal dimension should be captured into an integrated display [19].

Turoff et al. [45] developed design principles and specifications for a “Dynamic Emergency Response Management Information System” (DERMIS). Their premises and requirements concern systems for all types of crisis and emergency situations, such as systems dealing with fires, bombs, hazardous materials or transportation interruptions. The general requirements they set include that the system should be easy to learn after a specific training and adapted to the task requirements and roles of emergency personnel. The learning effort should be minimized by means of self-explanatory design. Functions regarding the communication process were omitted. From their research, they derive general design principles that are applicable for any emergency response system. Principles that are easily transferrable also to other systems are seen in the recommendations that the system directory should be hierarchical structured to provide easy access to the data and information in the system, that updates should be synchronized right away and that relevant information and data should be linked and presented as one unit of information [45].

Another approach to the design of human-machine interfaces is the idea to follow the naturalistic perceptions of an operator in emergencies as well as his decisions and actions. The authors of this concept used studies on emotional arousal (e.g. [10, 11, 30, 48]) and a model called direct perception-action coupling (DPAC) which merges direct perception and embodied cognition and has been developed to describe the available capacities of an operator under stress. Rahman et al. [37] hold the view that emergency controls should be compatible to the operator’s perceptual, cognitive and motoric abilities in a stress situation and adapt to his “immanent and remnant [mostly bodily] skills” [37]. As human-machine interaction is not seen as a process of symbolic or logical reasoning, principles of this approach are direct perception and embodied action. According to the authors, interaction takes place in an almost automatic and non-conscious way. To allow an automatic and intuitive handling of the system, human-machine interfaces should not be reduced to touch screens, push buttons or controls, which all look and feel similar but rather expedient controls that integrate the cognitive and motoric skills of an operator in a stressful situation. Comparable controls are simple, such as single-acting controls in firefighting or aerial systems [37].

2.3 Studies with Regard to Usability of Interfaces in Firefighting

Fires are worldwide feared threats, especially in warm areas, regularly leading to numerous fatalities, such as in the most recently Oakland warehouse and the Baghdad

Table 3. Studies with regard to usability of interfaces in firefighting.

	Area	Main recommendations
Jiang et al. (2004) [23]	Incident command system for firefighters	<ul style="list-style-type: none"> • Accountability of resources and personnel • Situation assessment through multiple information sources • No information overload • Resource allocation • Communication support • Minimum of direct interaction • Partial automation
Prasanna et al. (2013) [36]	Information and decision support system for fire emergency response	<ul style="list-style-type: none"> • Avoidance of information overload • Avoidance of work related stressors • Avoidance of attention tunnelling • Appropriate use of salience • Reduction of system complexities • Balance of automation • Low working memory requirements • Suitable use of mental models (measuring units, notations, abbreviations, colors)
Monares et al. (2011) [31]	Mobile collaborative application for firefighters	<ul style="list-style-type: none"> • Self-learning system by log file recording and processing • Individual information selection by information layers • Simple graphics, icons

hospital fires in 2016. The physical demands of firefighting [18] but also the psychophysical and psychological responses of firefighters [41] are known since the 1990s. However, research on computer-aided firefighting assistance considering the HCI began only later in the 2000s ([23], see Table 3), criticising the insufficient use of computers.

Using field studies, interviews and low-fidelity prototypes, Jiang et al. [23] developed an incident command system, which supports decisions of firefighters who have to take many factors and different sources of information into account. The evaluation showed the importance of redundancy and that direct interaction should be kept to a minimum. As a certain degree of automation helps the operator to handle challenging situations [23], their work resulted in software and hardware prototypes to assist spontaneous and opportunistic interactions for firefighters within a structure, introduced in Jiang et al. [24].

Prasanna et al. [36] investigated human-computer interfaces of fire emergency response systems in an empirical study. End users participated in their study on a fire operations system, using a software prototype for the evaluation. Prasanna et al. [36] followed the concept that emergency response information systems should support the operator's situation awareness (SA) to enable better decision making [26]. Prasanna et al. [36] found evidence for the applicability of enhancing SA over the three levels perception, comprehension and projection, introduced by Endsley [12], with perception

regarding the status, attributes and dynamics, comprehension, referring to a holistic, graphical aided information representation rather than isolated information sets of numbers, and projection regarding trends of situational parameters. Prasanna et al. [36] recommend to avoid information overload by layered information architectures and by outbalancing push- and pull-type information. Dashboards are viewed as the most helpful aid to avoid attention tunneling. They further warned that salience has to be used appropriately to help the operator focus, without misleading and confusing him. Also automation should be balanced well, because excessive use might push the user “out of the loop” [36].

“MobileMap” [31] is a mobile collaborative application designed to support fire-fighters in their decision-making and communication in urban emergency situations. The idea of the application was the possibility to analyze information after a fire in order to learn for the future. Therefore, the application recorded all interaction, which were then conceived to the analysis of the emergency and aided the decisions of the personnel afterwards. Features of importance for usability were among others pre-loaded maps of cities with several zoom levels, the salience of emergency specific depict points of interest (POIs, e.g. police stations and hospitals), the differentiation into several information layers, the extensive use of graphical elements with a variety of icons and the overall simple user interface design [31].

2.4 Studies with Regard to Usability of Interfaces in Nuclear Power Plants and Nuclear Emergencies

Unlike firefighting and emergency response systems, a general risk is inherent in nuclear systems, raising the question of the meaning of “Design for safety”, as pointed out by Boy and Schmitt [4]. Control room operators in power plants observe and work with highly complex systems where user interfaces need to offer effective, efficient and, above all, safe operation. Additionally, intelligent decision support system for nuclear emergencies are considered (see Table 4).

Carvalho et al. [5] aimed at the modernization of the ANGRA I power plant from a human factors perspective. The operations were observed by cognitive task analysis (CTA) to provide ideal decision support in the new system. In an emergency, the operators have to follow specific procedures, which are based on four major design aspects of technical accuracy, real-time receipt of the procedure task, easy comprehension and error elimination or at least reduction. The field studies and observations of the operators’ performance in simulators, together with heuristic and scenario-based evaluation, have shown major generalizable aspects for safety-critical displays. Attention should be paid to adequate font sizes and formats, especially in displays with complex information, color contrast and graphics in general to prevent information overload, especially in numeric information. Furthermore, it was recommended to match the size of icons to their function. In all, structured presentation informing the user about deviations of measured values from reference points, i.e. information should not be thought of and presented as data but in form of information supporting the operator in the task of the evaluation [5]. Carvalho et al. [6] added the critical point of media mixes requiring different cognitive resources.

Table 4. Studies with regard to usability of interfaces in nuclear power plants and nuclear emergencies.

	Area	Main recommendations
Carvalho et al. (2008) [5], Carvalho et al. (2011) [6]	Redesign of supervisory control operator interfaces in nuclear power production	<ul style="list-style-type: none"> • Adequate font sizes and formats • Appropriate color contrast • Use of graphics to avoid information overload • Size of icons matching the function • Structured presentation of information • Uniform media usage
Boy and Schmitt (2013) [4]	Control and management of nuclear power plant	<ul style="list-style-type: none"> • Determination of automation levels • Balancing of automation
Papamichail and French (2004) [34]	Decision support system for nuclear emergencies	<ul style="list-style-type: none"> • Support of understanding the system • Completeness of information • Format of output • Volume of output • Ease of use • Ease of learning • Information output at appropriate time • Flexibility/adaptability of the system • Performance (accomplishment of task) • Usefulness (performance of user)

A human-system integration approach, based on the resilience engineering introduced by Hollnagel et al. [20], was applied in the safety design of the control and management of a nuclear power plant in Boy and Schmitt [4], defining different automation levels. Their spectrum ranges from the computer offering no assistance, where the human has to decide on his own and perform every action himself, over the computer offering a certain amount of decision or action alternatives to autonomous decision-making and action. As automation leads to complexity but can also aid to reduce problems with procedure accumulation, it has to be well-balanced in the human-computer system [4].

The design and evaluation of the intelligent decision support system for nuclear emergencies by Papamichail and French [34] was human-centred. The aim was to help decision makers in the elaboration and ranking of alternatives [34]. General recommendations can be seen in the emphasis of facilitating the system understanding and the claim of completeness of the provided information. Easy system use and learnability are implied [34], building on definition of Bailey and Person [1]. In contrast to Hancock and Szalma [19], Frassl et al. [16] and Turoff et al. [45] demanding immediate information provision, Papamichail and French [34] pointed to times that are more suitable for information output than others are. Furthermore, the authors recommended that the system should be able to adapt or adjust to new conditions, demands or circumstances [34].

2.5 Empirical Studies with Regard to Usability in Other Fields

Some empirical studies with regard to usability in safety-critical contexts cannot be assigned to just one field, but are not to be neglected. Nilsson and Stølen [32], for instance, used results of their plurality of cross-sectional empirical studies with emergency response actors to define generalizable requirements for user interfaces, which they validated through observations and interviews in a training exercise. They consulted studies of avalanche rescuing, fire fighters, police and ambulance services (Table 5).

Based on these studies, it was concluded that user interfaces should present an “operational picture” [32], meaning a definition of the operational area, which is relevant during the emergency with information about all factors, like objects and persons involved in an incident [7], logging, which was based on the conclusions of Chittaro et al. [8] for a log of all occurrences during the emergency as well as accessing services for information [45], meaning that the information from different sources should be presented together. It was further pointed to resource management, i.e. presenting the available resources like personnel and equipment, taken up from Pottebaum et al. [35] based on description logistics and the work of Joshi [25] on map-based interfaces for emergency work. In line with Humayoun et al. [21], functionalities to support emergency personnel by providing plans and tasks explicit during the operations, which also reports the progress or fulfilment of the task, were recommended. For the design of the user interface, Nilsson and Stølen [32] further suggested designated hardware buttons, multimodal user interfaces [9] and augmented reality techniques [17].

Frassl et al. [16] developed a system for collecting, managing and distributing information in relief operations of disasters. In line with the spiral model of Boehm [2], the generated system has passed through several prototype iterations with feedback from assessment experts, logistics experts and other end-users. Resulting functional requirements for the behavior and provided services are partly generalizable. In terms of data management, Frassl et al. [16] recommended that information should be offered in report forms which support the assessment of the situation. The user interface should provide a situation map for visualization of all geo-referenced data similar. Regarding the device control, data from external sensors has to be processed automatically to relieve the user from configuration tasks. Another aspect is the instant synchronization

Table 5. Empirical studies with regard to usability in other fields

	Area	Main recommendations
Nilsson and Stølen (2011) [32]	Emergency response (avalanches, firefighting, police, ambulance)	<ul style="list-style-type: none"> • Operational picture • Log of occurrences • Easy information access • Resource management • Supply of plans and tasks • Designated hardware buttons, multimodal user interfaces • Augmented reality techniques
Frassl et al. (2010) [16], Frassl et al. (2012) [15]	Disaster relief operations	<ul style="list-style-type: none"> • Support for assessment of situation • Situation map • Automatic processing of data • Instant synchronization of information • Simplicity of the usage • Support of decision making by integrity of provided data • Autonomous work of system • Reliability in case of system failure • Interoperability (information exchange) • Frugality of the system

of information, also in situations of difficult network connectivity. Requirements to the characteristics for systems used in disaster management missions are seen in the simplicity of the usage and the implementation of only essential functions and configurations, additionally to be accompanied by help of the system. Decision-making should be supported by the integrity of the provided data. Therefore, the system should give details about how reliable the respective information is [16].

3 Discussion

A systematic, cross-sectional literature search was conducted for studies regarding usability and interface design in safety-critical contexts. In addition to the general usability literature and fundamental guidelines, five theoretical and literature based studies, three studies with regard to user interface design in firefighting, four studies with regard to nuclear power, a cross-sectional study integrating avalanche, firefighting,

ambulance and police work and a study with regard to disaster relief operations were identified and analyzed for generalizable usability recommendations for safety-critical systems. Five recommendations specific to or of enhanced importance in safety-critical systems were formulated:

Information processing efficiency is impaired under negative emotions, such as stress and anxiety ([13], see also Sect. 1.1). Keeping the required memory workload to a minimum was identified as a cross-sectional key-topic, demanded in theoretical work ([44], see Sect. 2.2) as well as in empirical work in firefighting ([36], see Sect. 2.3).

- **Recommendation 1:** Keep required memory workload to a minimum.

The use of simple graphics and simple structures was identified as a major topic in theoretical work [19] and interface design in power plants [5] and for firefighting [31].

- **Recommendation 2:** Use simple graphics and simple structures.

In line, the avoidance of information overload was found as an important aspect in interface design in firefighting [23, 36] as well as in power plants [5, 6].

- **Recommendation 3:** Avoid information overload.

Time pressure and time-criticality are inherent in safety-critical contexts. As a result, remarks on immediate and automatic updates of data were found in various literature, such as in theoretical work by Hancock and Szalma [19] and Turoff et al. [45] and empirical and prototype-based work for emergency response systems by Frassl et al. [15, 16]. However, Papamichail and French [34] pointed to the differences in timely suitability of information output. It is concluded that the provision of new data, or information as prepared data for easier information processing, should be bound to timely and relevance restriction.

- **Recommendation 4:** Provide important updated information as fast as possible.

Automation is often used to support the user by now. However, the “operator-out-of-the-loop” phenomenon, where the user overly relies on the system, is well known [12]. In order to keep the human in the loop, balancing the degree of automation was found to be of importance in firefighting [23, 36] and nuclear power control [4].

- **Recommendation 5:** Keep the operator in the loop by a suitable automation degree.

However, recommendations alone cannot ensure high usability standards in systems. Formative evaluations with end-users, if at all possible, and iterative system design are needed, as also generally established in ISO 9241 part 210 “Human-centred design processes for interactive systems” [22]. A very popular model is the spiral model by Boehm [2], applied for example by Frassl [15, 16] for disaster management tool design. Also regarding safety critical-systems, Boring et al. [3], for instance, reemphasized the importance of usability evaluations in the context of nuclear power plant control units as a formative method during the early design stages, not as a summative method at the end of a project.

Acknowledgments. The literature review was performed within the project DURCHBLICK, funded by the German Federal Ministry of Education and Research (BMBF) in the context of the national program “Research for Civil Security” and the call “Civil Security – Aspects and Measures of Coping with Terrorism”.

References

1. Bailey, J.E., Pearson, S.W.: Development of a tool for measuring and analyzing computer user satisfaction. *Manage. Sci.* **29**, 530–545 (1983)
2. Boehm, B.W.: A spiral model of software development and enhancement. *Computer* **21**(5), 61–72 (1988)
3. Boring, R.L., Joe, J.C., Ulrich, T.A., Lew, R.T.: Early-stage design and evaluation for nuclear power plant control room upgrades. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **58** (1), 1909–1913 (2014)
4. Boy, G.A., Schmitt, K.A.: Design for safety: a cognitive engineering approach to the control and management of nuclear power plants. *Ann. Nucl. Energy* **52**, 125–136 (2013)
5. Carvalho, P.V., dos Santos, I.L., Gomes, J.O., Borges, M.R., Guerlain, S.: Human factors approach for evaluation and redesign of human–system interfaces of a nuclear power plant simulator. *Displays* **29**(3), 273–284 (2008)
6. Carvalho, P.V., Gomes, J.O., Borges, M.R.: Human centered design for nuclear power plant control room modernization. In: *CEUR Proceedings 4th Workshop HCP Human Centered Processes*, 10–11 February, pp. 25–33. Federal University of Rio de Janeiro (2011)
7. Chen, R., Sharman, R., Rao, H.R., Upadhyaya, S.J.: Coordination in emergency response management. *Commun. ACM* **51**(5), 66–73 (2008)
8. Chittaro, L., Zuliani, F., Carchietti, E.: Mobile devices in emergency medical services: user evaluation of a PDA-based interface for ambulance run reporting. In: Löffler, J., Klann, M. (eds.) *Mobile Response 2007*. LNCS, vol. 4458, pp. 19–28. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-75668-2_3](https://doi.org/10.1007/978-3-540-75668-2_3)
9. Cohen, P.R., McGee, D.R.: Tangible multimodal interfaces for safety-critical applications. *Commun. ACM* **47**(1), 41–46 (2004)
10. Damasio, A.R.: *The Feelings of What Happens: Body and Emotion in the Making of Consciousness*. Hartcourt, New York (1999)
11. Driskell, J.E., Salas, E.: *Stress and Human Performance*. Lawrence Erlbaum Associates, Mahwah (1996)
12. 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)
13. Eysenck, M.W., Calvo, M.G.: Anxiety and performance: the processing efficiency theory. *Cognit. Emot.* **6**(6), 409–434 (1992)
14. Eysenck, M.W., Derakshan, N., Santos, R., Calvo, M.G.: Anxiety and cognitive performance: attentional control theory. *Emotion* **7**(2), 336–353 (2007)
15. Frassl, M., Lichtenstern, M., Angermann, M.: *Disaster Management Tool (DMT)-Usability Engineering, System Architecture and Field Experiments*. German Aerospace Center (DLR), Wessling (2012)
16. Frassl, M., Lichtenstern, M., Khider, M., Angermann, M.: Developing a system for information management in disaster relief-methodology and requirements. In: *Proceedings of the 7th International ISCRAM Conference*, Seattle, USA (2010)

17. Fröhlich, P., Simon, R., Kaufmann, C.: Adding space to location in mobile emergency response technologies. In: Löffler, J., Klann, M. (eds.) *Mobile Response 2007*. LNCS, vol. 4458, pp. 71–76. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-75668-2_8](https://doi.org/10.1007/978-3-540-75668-2_8)
18. Gledhill, N., Jamnik, V.K.: Characterization of the physical demands of firefighting. *Canadian journal of sport sciences* **17**(3), 207–213 (1992)
19. Hancock, P.A., Szalma, J.L.: Operator stress and display design. *Ergon. Des.* **11**(2), 13–18 (2003)
20. Hollnagel, E., Woods, D.D., Leveson, N.: *Resilience Engineering: Concepts and Precepts*. Ashgate, Aldershot (2006)
21. Humayoun, S.R., Catarci, T., Leoni, M., Marrellam, A., Mecella, M., Bortenschlager, M., Steinmann, R.: Designing mobile systems in highly dynamic scenarios: the WORKPAD methodology. *Knowl. Technol. Policy* **22**(1), 25–43 (2009)
22. ISO 9241:2006: Ergonomics of human-system interaction. Part 110: Dialogue principles and part 210: Human-centred design for interactive systems
23. Jiang, X., Chen, N.Y., Hong, J.I., Wang, K., Takayama, L., Landay, J.A.: Siren: context-aware computing for firefighting. In: Ferscha, A., Mattern, F. (eds.) *Pervasive 2004*. LNCS, vol. 3001, pp. 87–105. Springer, Heidelberg (2004). doi:[10.1007/978-3-540-24646-6_6](https://doi.org/10.1007/978-3-540-24646-6_6)
24. Jiang, X., Hong, J.I., Takayama, L.A., Landay, J.A.: Ubiquitous computing for firefighters: field studies and prototypes of large displays for incident command. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 679–686. ACM (2004)
25. Joshi, S.G.: Exploring map-based interfaces for mobile solutions in emergency work. Masters thesis, University of Oslo (2011)
26. Klann, M., Malizia, A., Chittaro, L., Cuevas, I.A., Levialdi, S.: HCI for Emergencies. In: *Proceedings of the 26th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI 2008)*, 5–10 April, Florence, Italy, pp. 3945–3948 (2008)
27. Lazarus, R.S.: *Emotion and Adaptation*. Oxford University Press, New York (1991)
28. Lazarus, R.S.: *Stress and Emotion: A New Synthesis*. Springer, New York (1999)
29. Lazarus, R.S., Folkman, S.: *Stress, Appraisal, and Coping*. Springer, New York (1984)
30. LeDoux, J.E.: *The Emotional Brain*. Simon and Schuster, New York (1996)
31. Monares, Á., Ochoa, S.F., Pino, J.A., Herskovic, V., Rodriguez-Covili, J., Neyem, A.: Mobile computing in urban emergency situations: improving the support to firefighters in the field. *Expert Syst. Appl.* **38**(2), 1255–1267 (2011)
32. Nilsson, E.G., Stølen, K.: Generic functionality in user interfaces for emergency response. In: *Proceedings of the 23rd Australian Computer-Human Interaction Conference*, pp. 233–242. ACM (2011)
33. Nielsen, J.: *Usability Engineering*. Morgan Kaufmann, Amsterdam (1994)
34. Papamichail, K.N., French, S.: Design and evaluation of an intelligent decision support system for nuclear emergencies. *Decis. Support Syst.* **41**(1), 84–111 (2004)
35. Pottebaum, J., Konstantopoulos, S., Koch, R., Paliouras, G.: SaR resource management based on description logics. In: Löffler, J., Klann, M. (eds.) *Mobile Response 2007*. LNCS, vol. 4458, pp. 61–70. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-75668-2_7](https://doi.org/10.1007/978-3-540-75668-2_7)
36. Prasanna, R., Yang, L., King, M.: Guidance for developing human–computer interfaces for supporting fire emergency response. *Risk Manage.* **15**(3), 155–179 (2013)
37. Rahman, M., Balakrishnan, G., Bergin, T.: Designing human–machine interfaces for naturalistic perceptions, decisions and actions occurring in emergency situations. *Theoret. Issues Ergon. Sci.* **13**(3), 358–379 (2012)
38. Sauro, J.: *Quantifying the User Experience: Practical Statistics for User Research*, 2nd edn. Morgan Kaufmann, Amsterdam (2016)
39. Sharp, H., Rogers, Y., Preece, J.: *Interaction design: beyond human-computer interaction*. Wiley, Chichester (2007)

40. Shneiderman, B.: *Designing the User Interface: Strategies for Effective Human Computer Interaction*. Addison-Wesley, Boston (1998)
41. Smith, D.L., Petruzzello, S.J., Kramer, J.M., Misner, J.E.: Physiological, psychophysical, and psychological responses of firefighters to firefighting training drills. *Aviat. Space Environ. Med.* **67**(11), 1063–1068 (1996)
42. Staal, M.A.: *Stress, Cognition, and Human Performance: A Literature review and Conceptual Framework*. Ames Research Center, Moflett Field (2004)
43. Starcke, K., Brand, M.: Decision making under stress: a selective review. *Neurosci. Biobehav. Rev.* **36**(4), 1228–1248 (2012)
44. Szalma, J.L., Hancock, P.A.: Task loading and stress in human-computer interaction: theoretical frameworks and mitigation strategies. In: Jacko, J.A. (ed.) *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*, pp. 115–132. CRC Press, Boca Raton (2008)
45. Turoff, M., Chumer, M., Van de Walle, B., Yao, X.: The design of a dynamic emergency response management information system (DERMIS). *JITTA: J. Inf. Technol. Theory Appl.* **5**(4), 1–3 (2004)
46. Welker, H.: *Handbook of Usability Engineering*. White Word Publications, Delhi (2012)
47. Wickens, C., Hollands, S., Banury, S., Parasuraman, R.: *Engineering Psychology*. Pearson Education Inc., Boston (2013)
48. Yerkes, R.M., Dodson, J.D.: The relation of strength of stimulus to rapidity of habit-formation. *J. Comp. Neurol. Psychol.* **18**, 459–482 (1908)

The Study of Presentation Characteristics of the Warning Information and Its Influence on User's Cognitive Process Based on Eye Tracking

Yun Lin, Chengqi Xue^(✉), Qi Guo, Jing Zhang, Ningyue Peng,
and Yafeng Niu

School of Mechanical Engineering,
Southeast University, Nanjing 211189, China
ipd_xcq@seu.edu.cn

Abstract. This research has adopted eye movement technique to study the influence of warning character on the information processing, which involves in two experiments. With our study, we have made our focus of research on the warning position, warning icon and warning border as a visual stimulus means. In our investigation, we have been keeping on with such commonly-made eye-movement recording parameters, such as the Fixation Count, the First Fixation and Duration by using an Eye Link II eye tracker, which is in a position to reflect the subject's attention and conversion of the attentions. What is more, we have done a single factor variance analysis (ANOVA) in hoping to work out the experimental data due to the kinds of eye movement parameters, and the following conclusions were drawn: (1) The warning position on the warning interface affects the machining process of the warning. When the warning is embedded in text, the warning is more noticeable and the perceived hazard level is higher. (2) Consistent with the results of the relevant studies on warnings on product labels, there are also icon effects on the warning interface, and the icon can improve the salience of the warning itself and the level of perceived danger. (3) There are also border effects on the warning interface. The appearance of the border makes the warning more significant and the perceived hazard level can be improved. The research results of this paper can also be adopted as the warning message design reference, which has had a great significance in improving the identification of warning message and reducing the rate of visual accidents on interface.

Keywords: Digital interface · Warning characteristics · Warning effectiveness · Eye tracking

1 Introduction

Warning sign is an important part of safety signs and widely used in industrial production, transportation, life and other fields in the past [1]. With the rapid development of technology and the Internet, more and more user interfaces are replaced by digital interfaces. These interfaces contain a large number of dynamic, complex information. Man identify and perceive danger information conveyed by warning information to

guide their behavior, so as to effectively avoid possible accidents. The process of the warning information has an effect on human behavior can be divided into four stages: discovery (attention stage), recognition (recognition stage), information judgment and decision (judgment stage), and compliance operation (behavior stage) [2]. In the whole process, the discovery stage is the prerequisite to comply with the operation, that is, the visual attention level of the warning information directly affects the human observance behavior of safety. Therefore, it is important to study the visual attention feature of the warning information, which can improve the recognition of the warning information and reduce the occurrence of the accident.

2 Background

2.1 Warning Information Processing Model

Wogalter and Laughery (1996) [3] have also applied the general information Communication Theory model to the field of warning research and proposed Human Information Processing, which considers warnings processing includes the following five cognitive processes: attention, understanding, attitudes and beliefs, Motivation, compliance behavior. Wogalter et al. (1999) [4] extended this information processing model, fitting communication components such as information sources, transmission channels, and receiver characteristics into the model, and proposed Communication-Human Information Processing model (C-HIP). Laughery (2006) [5] proposed an information communication model in the field of warning, and improved the C-HIP model. Laughery believed that the information communication model consists of four parts: the sender of the warning message, the warning message, the media, the recipient of the information, and discussed the influence factors of warning effectiveness from three aspects: the characteristics of the warning itself, the individual characteristics and the scene characteristics. Wu Xiaoli (2015) [6] proposed Error-Cognition Mapping from the perspective of the error factor and bring it into the complex information interface design. Figure 1 is a review and reorganization of warning information processing model.

2.2 Eye Movements and Warning Research

When individuals are browsing visual information, the visual information will be passed to the brain through visual system, the brain through the control of human eye movement to express interest in visual information, and this process is known as visual perception [7]. Individuals usually express their interest in visual objects or areas through initiative, frequent gaze. In cognitive psychology, the concept of gaze is called “selective attention”. Visual selective Attention use an information processing bottleneck mechanism, which allows only a small part of the information into the human eye to reach the short-term memory and visual attention area. Significant enough visual stimulus can be highlighted from the entire complex scene, and this significance has nothing to do with the purpose of observation, it occurs quickly in a bottom-up manner [8]. This mechanism of visual selection of attention can enable individuals to quickly locate objects or regions of interest

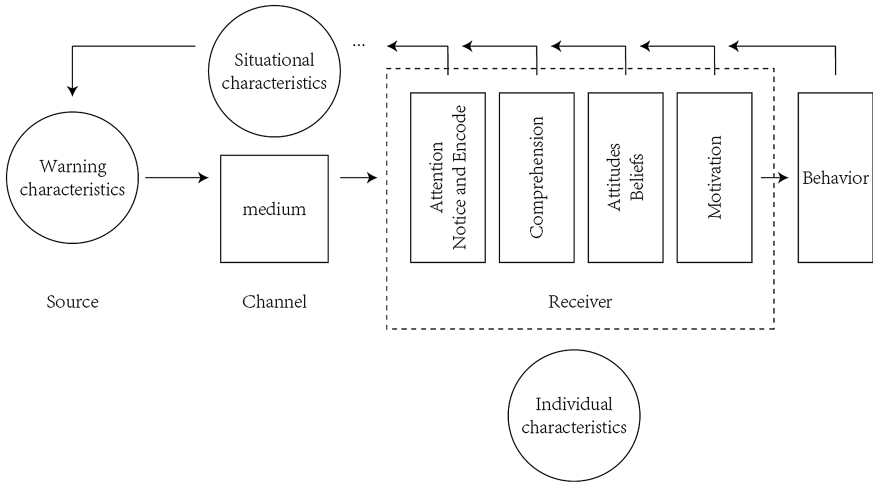


Fig. 1. The warning information processing model

in complex visual environments [9]. Test data of eye movement reflects individual attention and the conversion of attention. Therefore, the attention degree of visual stimuli can be obtained by collecting and analyzing the corresponding eye movement data.

3 Study 1: Effect of Warning Position and Warning Icon on the Warning Information Processing

Based on the previous research, experiment 1 will study the characteristics of the individual reading warning interface under different presentation characteristics of the warning using a 2-level factorial design. Variable 1: the warning position, including warning information embedded in the text and located at the bottom of the text. Variable 2: the warning icon, including warning information with icon and without icon.

3.1 Participates

A total of 14 male and 26 female graduate students aged between 23 and 27 years ($M = 23.8$, $SD = 2.1$) who used computer almost every day were recruited. All participants had normal or corrected vision without color blindness or color weakness. The participants would get a gift after the experiment.

3.2 Experimental Materials

The experiments were conducted in the ergonomics lab of Southeast University under normal lighting condition (about 300 lux). The stimuli were generated by one computer

with a 2 GHz Intel Core i7 processor. The computer was running under the Mac OS operating system. The display used was a 23.8-inch LCD monitor (Dell u2414 h). The graphics adapter was used at a resolution of 1920 × 1080 pixels and a frame rate of 60 Hz. The viewing distance used was 50 cm.

First of all, according to the principle that the experimental material should be similar to the real environment as much as possible to ensure the external validity and extensibility of experimental results, the experimental materials used in this study were selected from a number of control systems which include warning interfaces. The interfaces included a section of reading material, which contained warning information (Fig. 2). Contents of the reading material covered many areas, including driving, flight control, scheduling, organization and management. A total of 10 interfaces was selected.

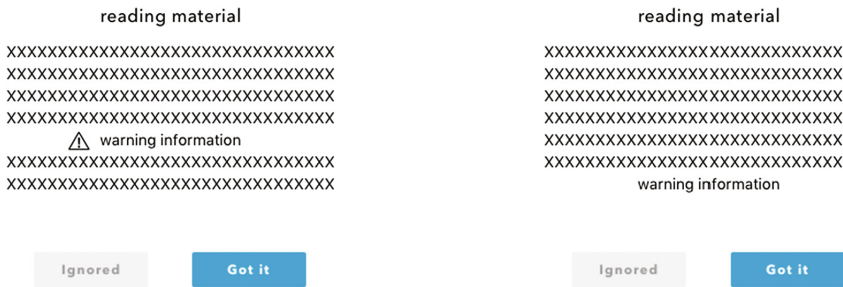


Fig. 2. Experimental material (embedded text, icon) & (bottom of text, no icon)

Next, some modifications were made to the raw material to ensure that when participants viewed each warning interface, the familiarity and reading time is roughly the same. All warning interfaces layouts used the same format—the outline paragraph structure. Each experimental material had approximately 3–4 paragraphs and there was only one warning message on each interface. Word number of each material was also roughly the same, being controlled between 158 and 172. In order to avoid different brands affected the participants reading the experimental material, the material’s brand was hidden. All factors of the warning information (fonts, spacing, etc.) were controlled to be consistent except for the warning position and warning icon. Experiment 1 was designed to examine the influence of warning position and warning icon on the warning information processing, adopting 2 (layout: embedded text, bottom of text) × 2(icon: yes, no) two factors within subjects design. The layout refers to the location of the warning information in the warning interface. “Embedded text” means that the warning information is presented near the text associated with the warning information, “bottom of the text” means the warning information is separated from the text and presented separately at the bottom of the page. The icon refers to whether a warning icon is displayed in the warning information, see Fig. 2. In this study, a triangle with an exclamation mark is used as a warning icon, see Fig. 3.

Finally, 2 warning interfaces were selected randomly from the 10 warning interfaces using as exercise materials and the remaining 8 warning interfaces were used as the formal experimental materials for the participants to read. In order to ensure that



Fig. 3. Warning icon

each participant observed each warning interface and read each alarm format, the content of the warning interface and the appearance order of the warning format were carried out (latin squares) balanced design.

3.3 Procedure

The task of the subject was to read the theme carefully and try to remember the contents of the warning interface. The participants were presented the 8 warning interfaces one by one. The “Space” button was used to turn the pages after each material had been read. Inform the participants that after the reading was complete, they need to fill out a paper questionnaire and determined the sentence in the questionnaire right or wrong according to the contents they had read. Before the start of the formal experiment, the participant first performed an exercise. Before the formal experiment began, the participants were given an exercise. Exercise was to read 2 warning interfaces continuously and judge four sentences which had been show to the participants right or wrong after the reading is complete.

In order to balance the reading material and warning feature sequence effect, the experiment will be randomly divided into 8 groups, so that each material can only be read once, participants can also read all the warning features. In order to ensure that the similarity between the experiment and the real environment, after reading all the materials, participants were presented an interference task to detect whether the warning information into the long-term memory before memory test.

4 Results

4.1 Memory Performance

The correct rate of the questionnaire part was compared, the results in Table 1.

Table 1. The mean and standard precision of the correct rate under different conditions

Position	Embedded text		Bottom of text	
	M	SD	M	SD
No	0.83	0.27	0.79	0.30
Yes	0.84	0.29	0.83	0.31

1. Main effects of warning position showed no significant ($F(1,39) = 0.291, P > 0.05$). Warning information embedded in text and at the bottom of text had no significant effects on the correct rate.
2. Main effects of warning icon showed no significant ($F(1,39) = 0.291, P > 0.05$). Having icon and no icon had no significant effects on the correct rate.
3. There was no significant interaction effect of warning position and warning icon ($F(1,39) = 0.053, P > 0.05$). It showed that there is no significant difference between warning position and warning icon in the correct rate.

4.2 Time of Entering the AOI

The time of entering the area of interest (AOI) when the participants read the warning interface was compared. The results are shown in Table 2.

Table 2. The mean and standard deviation of the time of entering the AOI under different conditions (s)

Position	Embedded text		Bottom of text	
	M	SD	M	SD
No	18.44	8.15	26.81	10.43
Yes	16.62	8.12	25.35	10.54

The analysis of variance (ANOVA) method was used to analyze the data. The results showed as follows:

1. Main effects of warning position showed significant ($F(1,30) = 61.422, P < 0.001$). Warning information embedded in text and at the bottom of text had significant effects on the participants' time of entering the AOI. Warning information embedded in text had a significantly shorter entry time than it at the bottom of the text.
2. Main effects of warning icon were significant ($F(1,30) = 4.298, P < 0.05$). Having icon and no icon had significant effects on the participants' time of entering the AOI. Having icon had a significantly shorter entry time than no icon.
3. There was no significant interaction effect of warning position and warning icon ($F(1,30) = 0.047, P < 0.05$). It showed that there is no significant difference between warning position and warning icon in the participants' time of entering the AOI.

4.3 The Number of Fixations in the AOI

The number of fixations in the AOI when the participants read the warning interface was compared. The results are shown in Table 3.

Table 3. The mean and standard deviation of the number of fixations in the AOI under different conditions (times)

Position	Embedded text		Bottom of text	
	M	SD	M	SD
No	14.63	5.85	12.21	4.31
Yes	17.74	6.16	12.03	5.17

The analysis of variance (ANOVA) method was used to analyze the data. The results showed as follows:

1. Main effects of warning position showed significant ($F(1,30) = 18.461, P < 0.01$). Warning information embedded in text and at the bottom of text had significant effects on the participants' time of entering the AOI. The number of fixations in embedded text was significantly greater than it at the bottom of the text.
2. Main effects of warning icon were significant ($F(1,30) = 7.869, P < 0.01$). Having icon and no icon had significant effects on the participants' time of entering the AOI. The number of fixations in having icon was significantly greater than it in no icon.
3. There was a significant interaction effect of warning position and warning icon ($F(1,30) = 6.799, P < 0.05$). Further analysis found that when the warning information was embedded in the text, there was a significant difference in the number of the participants' fixation in having warning icon and no icon. The number of fixations in having icon was significantly greater than it in no icon under this conditions. When the warning information was located at the bottom of the text, there was no significant difference in the number of the participants' fixation in having warning icon and no icon. Regardless of whether the warning information has an icon, the participants had significant differences in the number of fixations in embedded text and bottom of text. Warnings embedded in text have significantly more fixations than warnings at the bottom of text. see Fig. 4.

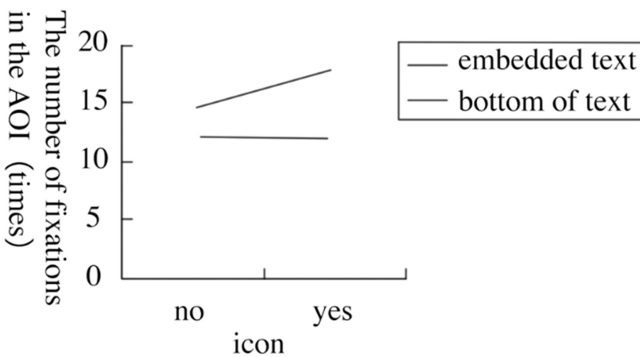


Fig. 4. The number of fixations in the AOI under different warning characteristics

4.4 AOI Total Residence Time

AOI total residence time when the participants read the warning interface was compared. The results are shown in Table 4.

Table 4. The mean and standard deviation of AOI total residence time under different conditions (s)

Position	Embedded text		Bottom of text	
	M	SD	M	SD
No	3.85	1.54	3.22	1.37
Yes	4.57	1.64	3.16	1.52

The analysis of variance (ANOVA) method was used to analyze the data. The results showed as follows:

1. Main effects of warning position showed significant ($F(1,30) = 15.080$, $P < 0.01$). Warning information embedded in text and at the bottom of text had significant effects on the participants' time of entering the AOI. The number of fixations in embedded text was significantly greater than it at the bottom of the text.
2. Main effects of warning icon were significant ($F(1,30) = 4.575$, $P < 0.01$). Having icon and no icon had significant effects on the participants' time of entering the AOI. The number of fixations in having icon was significantly greater than it in no icon.
3. There was a significant interaction effect of warning position and warning icon ($F(1,30) = 4.647$, $P < 0.05$). Further analysis found that when the warning information was embedded in the text, the participants had a significant difference in the AOI total residence time between the icon warning and no icon warning. The AOI total residence time for the icon warning was significantly longer than the no icon warning. When the warning information was located at the bottom of the text, there was no significant difference in the AOI total residence time between the icon warning and no icon warning. Regardless of whether the warning information has an icon, the participants' AOI total residence time had no significant differences in the warning position. See Fig. 5.

4.5 The Percentage of the AOI Residence Time Occupies Total Reading Time

The percentage of the AOI residence time occupies total reading time when the participants read the warning interface was compared. The results are shown in Table 5.

The analysis of variance (ANOVA) method was used to analyze the data. The results showed as follows:

1. Main effects of warning position showed significant ($F(1,30) = 19.685$, $P < 0.01$). Warning information embedded in text and at the bottom of text had significant effects on the percentage of the residence time occupies total reading time.

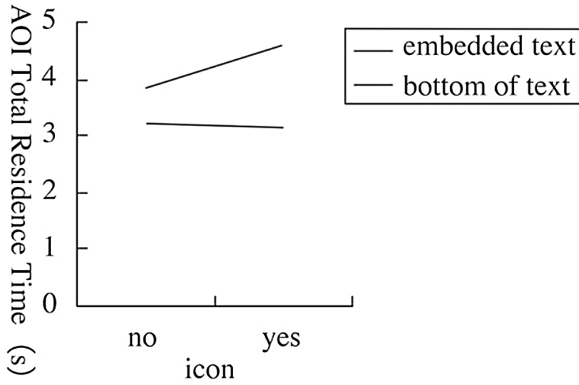


Fig. 5. The AOI total residence time under different warning characteristics

Table 5. The mean and standard deviation of the percentage of the AOI residence time occupies total reading time under different conditions (%)

Position	Embedded text		Bottom of text	
	M	SD	M	SD
No	3.59	1.11	2.97	1.37
Yes	4.16	1.76	3.26	1.37

The percentage of the embedded text residence time occupies total reading time was significantly more than the bottom of text residence time occupies total reading time.

2. Main effects of warning icon were significant ($F(1,30) = 4.285, P < 0.05$). Having icon and no icon had significant effects on the percentage of the embedded text residence time occupies total reading time. The percentage of having icon residence time occupies total reading time was significantly more than no icon residence time occupies total reading time.
3. There was no significant interaction effect of warning position and warning icon ($F(1,30) = 1.401, P > 0.05$). It showed that there is no significant difference between warning position and warning icon in the percentage of the AOI residence time occupies total reading time.

5 Discussion

Experiment 1 results showed that the position of warnings affected the warnings information processing. According to the Gestalt Psychology, individuals in life are perceived according to certain organizational rate experience. Continuity is a good organizing principle. The parts which are close to each other are more easily formed into the whole body for processing.

When the warning information was embedded in the text, the warning interface could be read more fluently, individuals could better perceive the specific content and point of the warning information. They could be aware of the danger would happen in which specific step of the operation and how it should be prevented directly and clearly. But when placed at the bottom of the text, the warning message has no explicit directivity, resulting in individual underestimating the importance of the warning, and the warning is not deep-level processing.

Similarly, a warning icon can also enhance the significance of the warning information and the danger degree of the warning message individual perceived. The presence of the icon in the warning made the participant notice the warning faster and longer. From the perspective of the choice mechanism of attention, the vast majority of warning information on warning interface was text, the emergence of icon stimulus could quickly attract people's attention, made people aware of the warning. At the same time, attention not only includes bottom-up processing, as well as top-down processing. The knowledge and experience individual already have as well as the expectations of the stimulus will also affect individual attention processing process. Warning icons are often associated with dangerous information, which has existed in individual long-term memory. When individual saw the warning icon again, the level of risk perception can be raised.

6 Study 2: Effect of Warning Position and Border on the Eye Gaze Warning

In this study, we investigated the characteristics of the individual reading warning interface under different presentation characteristics of the warning using a 2-level factorial design. Variable 1: the warning position, including warning information embedded in the text and located at the bottom of the text. Variable 2: the warning border, including warning information with border and without border.

6.1 Participates

A total of 15 male and 25 female graduate students aged between 23 and 27 years ($M = 23.3$, $SD = 2.6$) who used computer almost every day were recruited. All participants had normal or corrected vision without color blindness or color weakness. The participants would get a gift after the experiment.

6.2 Experimental Materials

The source of the experimental material is the same as the experiment 1. The content of the warning interface and the appearance order of the warning format were also carried out balanced design. Experiment 2 was designed to investigate the effect of warning position and border on the eye gaze warning, adopting 2 (layout: embedded text, bottom of text) \times 2 (border: yes, no) within subjects design. The border refers to whether a border is displayed in the warning information, see Fig. 4.

6.3 Procedure

Experimental procedure is the same as experiment 1.

7 Results

7.1 Memory Performance

The correct rate of the questionnaire part was compared, the results in Table 6.

Table 6. The mean and standard precision of the correct rate under different conditions

Position	Embedded text		Bottom of text	
	M	SD	M	SD
No	0.80	0.25	0.80	0.27
Yes	0.88	0.22	0.83	0.27

1. Main effects of warning position showed no significant ($F(1,28) = 0.375, P > 0.05$). Warning information embedded in text and at the bottom of text had no significant effects on the correct rate.
2. Main effects of warning border showed no significant ($F(1,28) = 0.191, P > 0.05$). Having icon and no icon had no significant effects on the correct rate.
3. There was no significant interaction effect of warning position and warning border ($F(1,28) = 0.415, P > 0.05$). It showed that there is no significant difference between warning position and warning border in the correct rate.

7.2 Time of Entering the AOI

The time of entering the area of interest (AOI) when the participants read the warning interface was compared. The results are shown in Table 7.

Table 7. The mean and standard deviation of the time of entering the AOI under different conditions (s)

Position	Embedded text		Bottom of text	
	M	SD	M	SD
No	17.62	6.38	24.99	7.23
Yes	16.04	7.08	23.57	6.57

The analysis of variance (ANOVA) method was used to analyze the data. The results showed as follows:

1. Main effects of warning position showed significant ($F(1,28) = 51.642, P < 0.001$). Warning information embedded in text and at the bottom of text had significant effects on the participants' time of entering the AOI. Warning information embedded in text had a significantly shorter entry time than it at the bottom of the text.
2. Main effects of warning border were significant ($F(1,28) = 4.335, P < 0.05$). Having border and no border had significant effects on the participants' time of entering the AOI. Having border had a significantly shorter entry time than no border.
3. There was no significant interaction effect of warning position and warning border ($F(1,30) = 0.013, P < 0.05$). It showed that there is no significant difference between warning position and warning border in the participants' time of entering the AOI.

7.3 The Number of Fixations in the AOI

The number of fixations in the AOI when the participants read the warning interface was compared. The results are shown in Table 8.

Table 8. The mean and standard deviation of the number of fixations in the AOI under different conditions (times)

Position	Embedded text		Bottom of text	
	M	SD	M	SD
Border				
No	14.07	4.47	11.29	5.36
Yes	16.02	6.08	12.74	5.12

The analysis of variance (ANOVA) method was used to analyze the data. The results showed as follows:

1. Main effects of warning position showed significant ($F(1,28) = 11.114, P < 0.01$). Warning information embedded in text and at the bottom of text had significant effects on the participants' time of entering the AOI. The number of fixations in embedded text was significantly greater than it at the bottom of the text.
2. Main effects of warning border were significant ($F(1,28) = 11.663, P < 0.01$). Having border and no border had significant effects on the participants' time of entering the AOI. The number of fixations in having border was significantly greater than it in no border.
3. There was a significant interaction effect of warning position and warning icon ($F(1,28) = 0.090, P > 0.05$). It showed that there is no significant difference between warning position and warning border in the number of fixations in the AOI.

7.4 AOI Total Residence Time

AOI total residence time when the participants read the warning interface was compared. The results are shown in Table 9.

Table 9. The mean and standard deviation of AOI total residence time under different conditions (s)

Position	Embedded text		Bottom of text	
	M	SD	M	SD
No	3.59	1.11	2.97	1.37
Yes	4.16	1.76	3.26	1.37

The analysis of variance (ANOVA) method was used to analyze the data. The results showed as follows:

1. Main effects of warning position showed significant ($F(1,28) = 9.441, P < 0.01$). Warning information embedded in text and at the bottom of text had significant effects on the participants' time of entering the AOI. The number of fixations in embedded text was significantly greater than it at the bottom of the text.
2. Main effects of warning border were significant ($F(1,28) = 8.793, P < 0.01$). Having border and no border had significant effects on the participants' time of entering the AOI. The number of fixations in having border was significantly greater than it in no border.
3. There was a significant interaction effect of warning position and warning border ($F(1,28) = 0.442 P > 0.05$). It showed that there is no significant difference between warning position and warning border in the AOI total residence time.

7.5 The Percentage of the AOI Residence Time Occupies Total Reading Time

The percentage of the AOI residence time occupies total reading time when the participants read the warning interface was compared. The results are shown in Table 10.

Table 10. The mean and standard deviation of the percentage of the AOI residence time occupies total reading time under different conditions (%)

Position	Embedded text		Bottom of text	
	M	SD	M	SD
No	3.59	1.11	2.97	1.37
Yes	4.16	1.76	3.26	1.37

The analysis of variance (ANOVA) method was used to analyze the data. The results showed as follows:

1. Main effects of warning position showed significant ($F(1,28) = 14.586, P < 0.01$). Warning information embedded in text and at the bottom of text had significant effects on the percentage of the residence time occupies total reading time. The percentage of the embedded text residence time occupies total reading time was significantly more than the bottom of text residence time occupies total reading time.
2. Main effects of warning border were significant ($F(1,28) = 5.233, P < 0.05$). Having border and no border had significant effects on the percentage of the embedded text residence time occupies total reading time. The percentage of having border residence time occupies total reading time was significantly more than no border residence time occupies total reading time.
3. There was no significant interaction effect of warning position and warning border ($F(1,28) = 0.794, P > 0.05$). It showed that there is no significant difference between warning position and warning border in the percentage of the AOI residence time occupies total reading time.

8 Discussion

From the results of Experiment 2, it can be deduced that the warning information and the border can attract the individual attention quickly, and last the attention to the warning information longer. It is worth noting that the results of the two experiment showed that participates' memory performance was no significant difference, which is inconsistent with previous studies. It may be due to the fact that the average reading time of the warnings in this study is long, which leads to no difference in memory scores. The three features of position, icon and border may only affect the phase of perceiving and understanding in the warning information processing, while no effect on the subsequent long-term memory stage. It is also possible that, because there are few experimental materials in this study, the subjects only need to read the eight warning interfaces without time limit, and the memory score test adopts the form of recognition, the task is simple, memory performance no difference.

9 Conclusion

In this study, two experiments were conducted to investigate the effect of warning features on the warning process, and the following conclusions were drawn:

1. The warning position on the warning interface affects the machining process of the warning. When the warning is embedded in text, the warning is more noticeable and the perceived hazard level is higher [6].
2. Consistent with the results of the relevant studies on warnings on product labels, there are also icon effects on the warning interface, and the icon can improve the salience of the warning itself and the level of perceived danger.

3. There are also border effects on the warning interface. The appearance of the border makes the warning more significant and the perceived hazard level can be improved.

Some suggestions for the warning interface designer:

The warning information should be presented in the description text. The study found that when the warning is embedded in the text, the warning reading time is significantly longer, providing greater possibilities for the following individual observance behavior of safety. At the same time, warnings on the warning interface should use features such as borders and icons to improve the significance of warnings.

Acknowledgments. This work was supported by the National Nature Science Foundation of China (Grant No. 71471037, 71271053).

References

1. Guoqing, N., Caicai, C., Kun, Z.: An eye movement study of auxiliary words effect to safety signs' recognition. *J. Henan Polytech. Univ: Nat. Sci.* **33**(4), 410–415 (2014)
2. Yicheng, H., Xiaohong, Z., Liang, W.: Evaluating effectiveness of safety signs on building site. *China Saf. Sci. J.* **22**(8), 37–42 (2012)
3. Wogalter, M.S., Laughery, K.R.: WARNING! Sign and Label Effectiveness. *Cwrr. Dir. Psychol. Sci.* **5**(2), 35–37 (1996)
4. Wogalter, M.S., DeJoy, D.M., Laughery, K.R.: *Warnings and Risk Communication*. Taylor and Francis, Philadelphia (1999)
5. Wogalter, M.S., Vigilante J.W.J.: Attention Switch and Maintenance. *Handbook of Warnings*, pp. 245–265. Lawrence Erlbaum Associates, Mahwah (2006)
6. Xiaoli, W.: *Complex Information Task Interface Error-cognitive Mechanism*. Southeast University (2015)
7. Mingru, H., Zheming, S., Yongjian, L.: Increasing availability of knowledge map based on eye movement experiment of cognitive load. *Chin. J. Manage.* **9**(5), 753–757 (2012)
8. Yanlan, S., Rui, Z., Cheng, Z., et al.: Visual attention based image classification. *J. Image Graph.* **10**, 1886–1889 (2008)
9. Nong, S., Zhenglong, L., Tianxu, Z.: Applications of human visual attention mechanism in object detection. *Infrared Laser Eng.* **33**(1), 38–42 (2004)

Cognitive Task Analysis for Interface Designs to Assist Medical Engineers in Hemodialysis Machine Troubleshooting

Yoshitaka Maeda¹(✉), Satoshi Suzuki², and Akinori Komatsubara¹

¹ Waseda University, Tokyo, Japan
y-maeda@aoni.waseda.jp

² Kanagawa Institute of Technology, Kanagawa, Japan

Abstract. With the aim of designing an interface that supports troubleshooting of a dialysis machine, a medical-engineer (ME) cognitive task analysis was conducted in this study, with the error messages currently provided by a hemodialysis machine also being analyzed and evaluated. First, we developed the “error-message mechanism diagram” for the given problem, indicating the relationship between the error message and the notifying conditions of this message (corresponding to the candidate for the cause of the problem). Next, we developed the “cognitive task flow diagram,” which shows the cause candidates generated by the ME until the source of the problem was detected. This diagram also clarifies the manner in which the ME verifies the cause candidates and the information or knowledge employed by the ME. Then, for the given problem, we compared the cognitive task flow diagram of an ME who successfully detected the problem cause and corresponding error-message mechanism diagram to evaluate the efficacy of the error messages currently provided by the device.

Keywords: Troubleshooting · Cognitive task analysis · Medical equipment and hemodialysis

1 Introduction

Hemodialysis is a medical treatment in which moisture and waste matter accumulated in the patients’ body because of renal failure are removed by the extracorporeal circulation of blood. Patients must receive treatments for approximately four hours, three times a week, at dialysis hospitals. If the renal function is not improved, waste materials drained from their kidneys accumulates in their body. Until December 2015, there were more than 320,000 dialysis patients in Japan; this number is on an increasing trend as the number of patients suffering from diabetic nephropathy increases, which is the cause of dialysis initiation [1].

The dialysis device (machine) used in such a hemodialysis treatment is indispensable for dialysis medical safety and quality improvement because it can circulate the patient’s blood extracorporeally at a constant speed, partly proceed to remove waste matter and moisture inside the body, and automatically detect abnormalities in the patient’s physical condition. For this reason, medical engineers (MEs) responsible for maintenance and management of machine are obliged to perform periodic/daily

inspections (inspection before and after treatment) of the machine, and the contents of inspections are bound by the guidelines of Japan Association for Clinical Engineers [2]. The machine operates by separating liquids, such as the dialytic fluid from the blood, within the internal circulatory system. This principle is the same for dialysis machines in other countries [3]. Further, a typical dialysis machine used in many hospitals in Japan has a self-checking function that can detect problems such as internal leaking or pressure abnormalities. If a problem occurs, the machine displays error messages on its monitor to indicate an irregularity, for example, “Pump voltage abnormality.” However, the cause of an irregularity and the involved machine parts are not identified in these messages (Fig. 1).



Fig. 1. Dialysis machine

The characteristics of such dialysis machines are similar to those of industrial plants. The Engineering Equipment and Material Users Association (EEMUA), which pertains to plant-related industry, has stated that a desirable error message should provide “detailed information of the causes of trouble, which can help in troubleshooting,” in their guidelines [4]. However, from a cost perspective, it is unfeasible to detect problematic parts automatically by attaching sensors to all the parts in a hemodialysis machine. Therefore, MEs responsible for the maintenance of these devices must detect the cause of a problem by gathering information on the internal status of the machine (e.g., the pressure or voltage) and by monitoring the interior of the machine, based on error messages. As the machine is always used for hemodialysis treatment, the speed and accuracy with which MEs can identify the causes of trouble and restore can greatly affect the patient safety and treatment quality.

Further, even in a general device (not only a dialysis machine), it is common for users to perform troubleshooting. Therefore, the interface must be designed to instruct the operation so that the user can deal with the trouble promptly and appropriately, and

to present the cause candidates in an easy-to-understand manner is important and is a major problem [5, 6].

In the current situation surrounding such troubleshooting support, to troubleshoot efficiently and effectively, it is important to detail the manner in which the error messages indicate the machine irregularities to the user.

With the aim of designing an interface that supports machine troubleshooting, an ME cognitive task analysis was conducted in this study, along with the analysis and evaluation of the error messages currently provided by a hemodialysis machine.

2 Method

2.1 Design

In this study, we conducted a cognitive task analysis of MEs who can efficiently and effectively identify the cause of trouble. We then developed a “behavioral model,” which shows the cause candidates generated by the ME until the source of the problem is detected. By designing the interface along this model, the contents of an alarm system is expected to be displayed, as required by MEs, on the machine monitor.

Next, we developed the “error-message mechanism diagram” for a given problem, indicating the relationship between the error message and the notifying conditions of this message (corresponding to the candidate for the cause of the problem).

This error message mechanism diagram is said to clearly show the current state of user support on the machine regarding cause identification. Therefore, we compared the “behavioral model” with the corresponding error-message mechanism diagram to evaluate the efficacy of the error messages currently provided by the device.

In the hospital, where the study was conducted, approximately 100 patients receive dialysis treatment in 52 beds daily.

2.2 MEs’ Behavioral Model During Troubleshooting

Five types of issues. One expert ME (with 13 years of experience) made five types of trouble-triggering mechanism in the dialysis machine (Nikkiso Co., DCS-27). Table 1 shows the contents of the trouble and error message displayed on the machine monitor.

Recording ME Behavior during Troubleshooting. Seven MEs with 4–25 years of experience performed troubleshooting tasks on a dialysis machine with five types of problems (Table 1). In this task, MEs performed routine inspections by using the self-diagnosis function of the machine. Specifically, as warnings related to each task were sequentially displayed on the monitor during self-diagnosis, MEs were asked to search for the cause of each problem, and they exchanged parts of the machine as necessary. Although no time limit was set for each task, we allowed the MEs to abandon the cause specification through a declaration.

Table 1. Summary of five types of problems.

Problem No.1	
▪ mechanism:	loosen the power supply connector of the deaeration pump in the machine
▪ problem:	A pump does not operate
▪ error message:	De-aeration pump is locked.
Problem No.2	
▪ mechanism:	Jam-pack one of two air filters adjacent to SV41
▪ problem:	SV41 cannot take in air
▪ error message:	SV41 test failure
Problem No.3	
▪ mechanism:	Damage a flow quantity adjustment valve in the deaeration pump fluid entrance side of the pump
▪ problem:	Fail to keep a fixed flow quantity
▪ error message:	pump cell No.1 is closed (the absorbing side)
Problem No.4	
▪ mechanism:	prevents the impeller in the pressurization pump from turning
▪ problem:	Fail to keep negative pressure in the plumbing
▪ error message:	A test of the decompression is unacceptable
Problem No.5	
▪ mechanism:	Scrach Diaphragm of back pressure valve
▪ problem:	Because machine cannot keep a fixed pressure in the plumbing, an plumbing leak test is not conducted appropriately
▪ error message:	The self-check result for the plumbing leakage is unacceptable

In this study, eye-tracking systems (EMR; EMR-8 manufactured by Nac Co.) and wearable cameras were attached to MEs. We then recorded gaze point at the time of cause identification (Fig. 2). In addition, we conducted the Virball Protocol Law (i.e., we recorded the MEs' free-standing voice about "what they were thinking at the moment," "what kind of cause candidate can be considered," and "what kind of intention they were gazing at"), and recorded behavior by using a video camera.

Development of ME Behavioral Model. Based on the eye-tracking records, cognitive task flow diagrams were depicted for each problem. They show the information and knowledge employed by the ME, the candidate cause generated by the ME, and the procedure used to verify the candidate cause.



Fig. 2. Video image (bottom right: eye camera image)

2.3 Development of Error-Message Mechanism Diagrams

Based on the machine manufacturer's instruction manual and alarm list, error-message mechanism diagrams were created for each problem, and show the candidate cause of trouble assumed by the machine manufacturer.

3 Results

3.1 Results of ME Behavioral Model/Error-Message Mechanism Diagrams

Figures 3, 4, and 5 show the error-message mechanism diagrams for Problem Nos. 2, 3, and 5, respectively, in Table 1 and the behavioral model of the ME who succeeded/failed in detecting the source of the problem within the shortest period.

The top row of each figure presents the error-message mechanism diagrams, showing the message contents and candidate causes indicated by the messages.

The lower row of each figure represents the ME behavioral model, showing the information and knowledge employed by the ME, the candidate cause generated by the ME, and the procedure used to verify the candidate cause. These items are shown in chronological order from left to right.

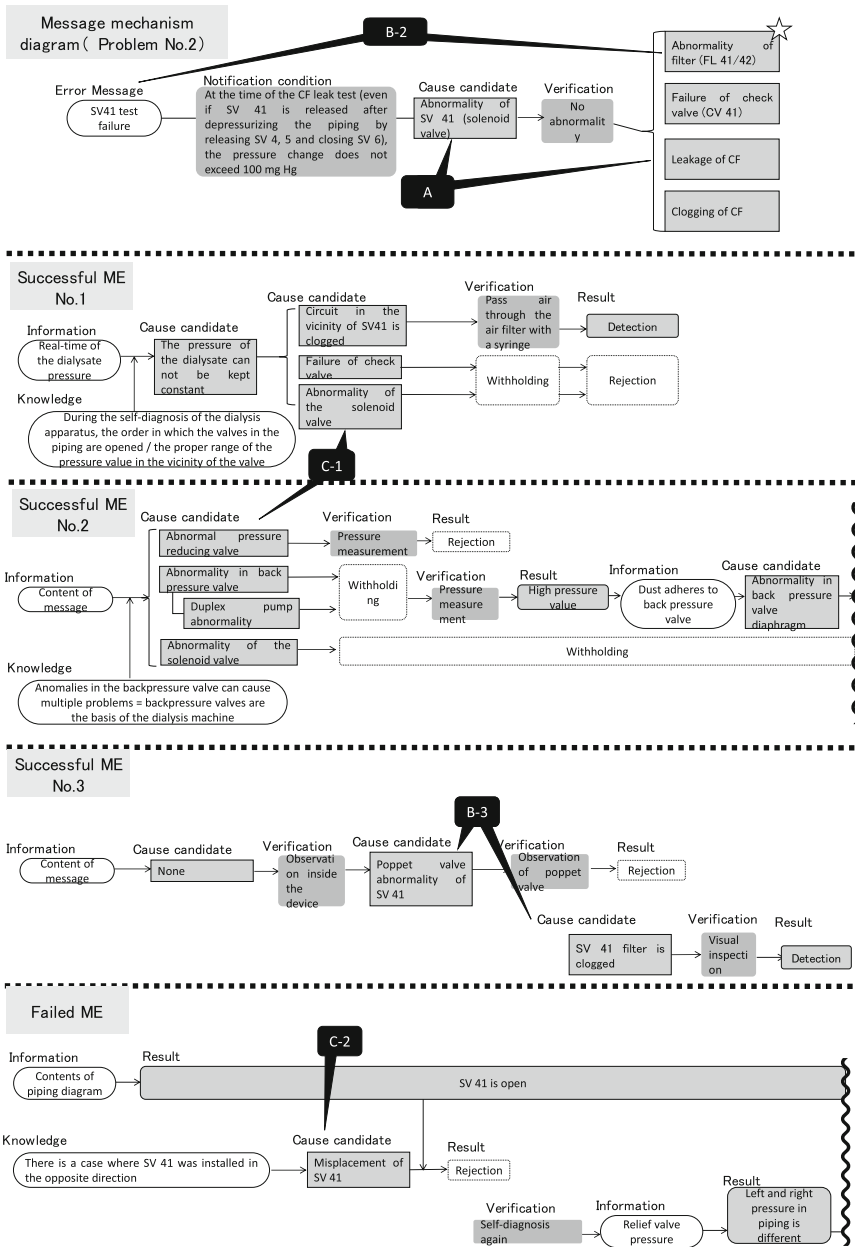


Fig. 3. Error-message mechanism diagram and behavioral model for Problem No. 2.

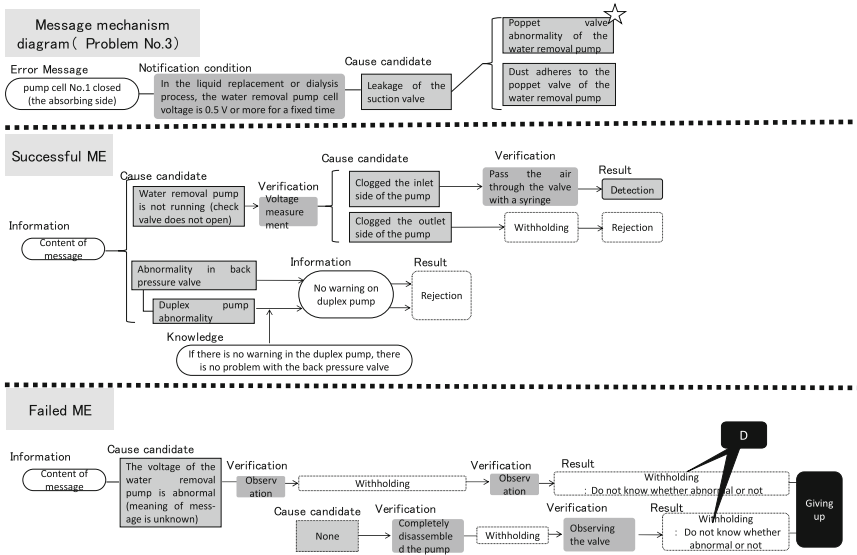


Fig. 4. Error-message mechanism diagram and behavioral model for Problem No. 3.

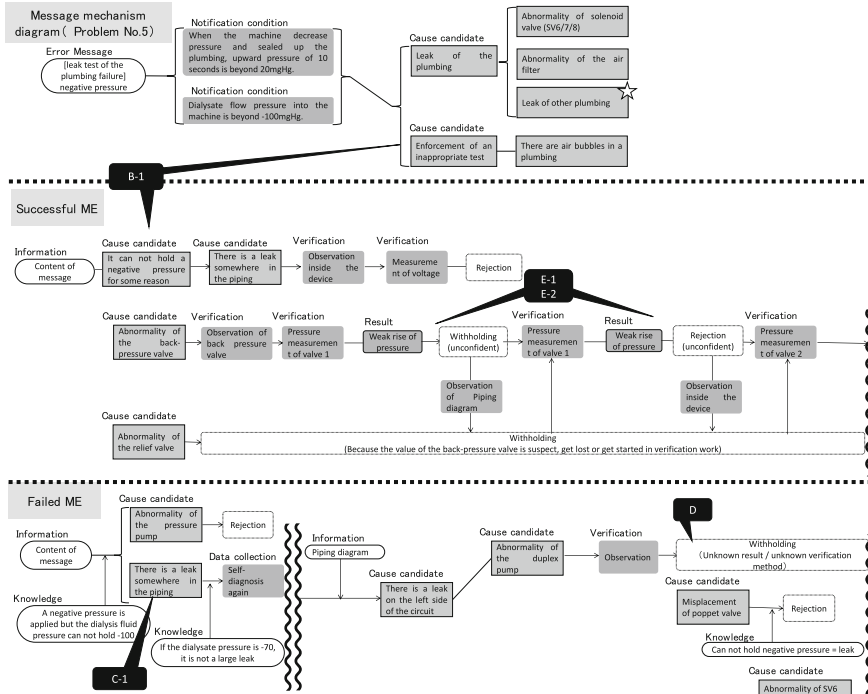


Fig. 5. Error-message mechanism diagram and behavioral model for Problem No. 5.

3.2 Comparison Result of Error-Message Mechanism Diagram and Behavioral Model

The comparison of the error-message mechanism diagram and behavioral model of MEs in each problem yields the following characteristics regarding the interface design related to the apparatus alarm.

Interpretation of MEs Regarding the Implication of the Error Message. In Problem No.1, the list of candidate causes was generally consistent with those of the error-message mechanism diagram. This could be because the ME recalled one irregular condition “pump is locked” based on the error message “De-aeration pump is locked.”

In contrast, in Problem Nos. 2, 4, and 5, the list of candidate causes generated by the ME was more extensive than the candidate causes presented in the error-message mechanism diagram (Fig. 3A). It is assumed that an ME considered two possible irregularities: “the self-check was not conducted appropriately” and “the self-check was conducted appropriately but there is a leak in the plumbing,” based on the error message: “The self-check result for the plumbing leakage is unacceptable.”

Moreover, with Problem No. 3, the MEs could not understand the meaning of the error message “pump cell No. 1 is closed (the absorbing side)” or could not generate the cause candidate (in Fig. 4: error-message mechanism diagram) assumed by the machine maker. Therefore, it seemed that this message was not able to support the cause identification of the MEs.

Furthermore, with the error message “A test of the decompression is unacceptable” in Problem No. 4, it was found that the cause candidate cannot be generated because the MEs cannot understand the type of mechanism used to conduct the test.

Support on the Verification Order of Multiple Cause Candidates. For each problem, the device monitor does not show the order in which candidates of the multiple-cause candidates in the alarm mechanism diagram should be verified (which candidate is the probable cause of high probability, etc.). Therefore, the verification order of the cause candidates varied among the MEs (Fig. 5: B-1).

Moreover, in Problem No. 2, despite the assumption by the machine maker that the defect of the machine part “SV 41” is the most probable cause candidate, it is not clearly displayed on the monitor (Fig. 3: B-2).

Furthermore, in Problem No. 2, the word “SV 41” is used as an error message even though the machine manufacturer assumes cause candidates for parts other than SV 41 (in the error-message mechanism diagram of Fig. 4); thus, some MEs only generated candidate causes of SV 41 (Fig. 3: B-3).

Number and Range of Cause Candidates indicated by Error Message. In Problem Nos. 2, 4, and 5, the cause candidates indicated by the error message (e.g., piping leakage) are in a wide range in the machine; thus, the cause candidates generated by the MEs are very large (Figs. 3 and 5: C-1). Therefore, even among the MEs who succeeded in identifying the cause, the cause candidates generated vary.

On the contrary, in Problem Nos. 2 and 4, the machine manufacturer did not assume the cause candidate generated by the MEs (in Fig. 3: error-message mechanism diagram). The content was a cause of trouble that could occur because of a mistake in setting up the machine in advance by a ME, for example, “valve mounting error” (Fig. 3: C-2).

Support for Verification of Cause Candidates. For the cause candidate generated by the MEs, the dialysis machine does not show the verification method for whether it is the true cause. In addition, the normal state of each part of the machine is not shown (Figs. 4 and 5D). Therefore, the MEs did not know the verification method, or made a mistake in the verification, that is, the machine does not appropriately support the narrowing down of the cause candidates.

Display of Pressure Value and Other Factors in the Machine Monitor. When generating a cause candidate, MEs often used a piping diagram (a simplified diagram of piping showing parameters, such as pressure values, in piping) displayed on the machine monitor (Fig. 6). In this piping diagram, the operating state of the pump was indicated in “flashing red”; however, in Problem No. 4, there was a case in which the ME mistook it as a troubleshooting site. In contrast, in this piping diagram, only the pressure value of the piping is shown, and it does not indicate in which area the pressure value is abnormal; thus, the machine cannot support verification of the cause candidate.



Fig. 6. Piping diagram.

Further, in addition to the piping diagram, the pressure and voltage values in the machine can be confirmed but they show the real-time and maximum pressure values in a certain period. Therefore, it is difficult for the MEs to verify the cause candidate by the transition of the pressure value (Fig. 5: E-1). In addition, the device did not show the transition of the pressure value in the normal state; this was a reference for judging whether the verification is defective (Fig. 5: E-2).

4 Discussion

4.1 Interface of the Dialysis Machine

EEMUA lists eight requirements for a desirable error message in a plant, such as a chemical factory, that has characteristics similar to a dialysis machine (Table 2) [4]. Therefore, in this section, we discuss the features of the alarm interface in the dialysis machine; this was clarified by comparing the error-message mechanism diagram and behavioral model of the MEs from the viewpoint of these eight requirements. Considering “Timely,” of the dialysis equipment, an error message was announced immediately after the occurrence of trouble, and the timing is assumed to be appropriate in troubleshooting by the MEs; thus, it was excluded from consideration.

Table 2. Eight requirements for a desirable error message in a plant (EEMUA) [4]

Requirement	Summary
Relevant	For each alarm, the corresponding operation after the occurrence of the alarm is indicated
Understandable	For each alarm, a message is shown to make it easier to understand the present situation, the expected future situation, and the corresponding operation to be taken
Unique	Each alarm is designed to notify one trouble
Prioritized	Each alarm is designed to provide a message indicating the priority of the corresponding operation
Diagnostic	For each alarm, information for diagnosing the current situation, the expected future situation, and the source of the trouble is provided
Advisory	Each alarm provides a guide for the operator to deal with the trouble
Focusing	Each alarm is designed so that an operator notices it quickly and reliably
Timely	Each alarm is designed to occur at an appropriate timing

Relevant/Understandable. As specific machine-part names were provided in the error message, many MEs did not generate cause candidates related to other parts of machine. It is important to show the error message in detail so that the corresponding operation after the alarm notification can be understood by MEs. However, from the earlier tendency, if a specific part name is included in the message, there is a possibility that other parts will not be handled as a cause of trouble by the MEs.

In addition, for an error message that the MEs cannot understand, for example, the alarm message “pump cell No. 1 in the water removal pump is closed (the absorbing side)” in Problem No. 3, the MEs have to generate a cause candidate by relying on the part name (in this case, the water removal pump) included in the message. Therefore, this message can be meaningless or obstructive for the cause identification by the MEs.

Unique. Originally, it is desirable to uniquely express one irregular device state, such as “De-aeration pump is locked” in the error message. However, with the accuracy of the trouble detection sensor attached to the current dialysis machine, it is difficult to

obtain the error-message uniqueness by matching the irregular state and message on a one-to-one basis. In other words, one error message might have multiple meanings.

In contrast, in the messages of Problem Nos. 2, 4, and 5 (such as “A test of the decompression is unacceptable”), the MEs may understand only one of the plurality of meanings (such as “The test was not successful” and “The test went well but failed”). In other words, there is a possibility that it will hinder the generation of the cause candidate by the MEs.

Moreover, when the trouble area indicated by the error message extends to the entire machine, the cause candidates generated by the MEs were in excess, thus tending to be difficult to verify.

Prioritized/Diagnostic/Advisory. For all problems, the machine did not show how to narrow down the cause candidates generated by the MEs (method of verifying whether it is a true cause), and the verification order of the multiple cause candidates is not clear. In addition, the machine did not show the verification criteria for the cause candidates, such as the appearance of the machine parts in the normal state and the value of the pressure in the machine. As a result, the MEs failed to conclude/reject the cause candidate, abandoned it, or misdiagnosed it; therefore, it is considered that these information are important for prompt and accurate cause identification.

In addition, with some troubleshooting, it was determined that MEs may verify the cause candidate by the trend of the machine’s pressure value. However, the present machine only shows real-time and maximum pressure values during a certain period, and it seemed that the MEs found it difficult to read the tendency of the pressure. Therefore, it seems that machine manufacturers must display data, such as pressure values, on the device monitor by using a display method that conforms to the method of verifying the cause candidate by the MEs.

Focusing. In the piping diagram used by almost all the MEs during the generation and verification of the cause candidate, in one case, the ME mistook the “flashing red,” indicating the operation state of the pump, as the trouble cause place. As described earlier, it is contemplated that a high-focus level display in the dialysis machine monitor should be avoided because there may be misunderstood as a warning indication at troubleshooting.

4.2 Characteristics of MEs Who Succeeded/Failed to Perform Troubleshooting

By comparing the behavior models of MEs who succeeded and failed in troubleshooting, we discuss the characteristics of the MEs as users of the dialysis machine.

Generation of Cause Candidate. The following tendencies were observed regarding the generation of cause candidates for MEs who failed troubleshooting.

- Generation of cause candidates biased toward one part or display (e.g., piping diagram).
- As the MEs never experienced a trouble in the contents of an error message, they cannot generate a cause candidate or can only generate a cause candidate within their individual experience.
- Generation of cause candidates unrelated to the error message (forcibly link candidates to the troubles experienced in the past).

Owing to these tendencies, it is considered that the experience of MEs in determining the trouble greatly affects the generation result of the cause candidate.

Verification Order of Cause Candidates. Among the MEs who generated multiple cause candidates, the MEs who failed troubleshooting conducted the verification in a random order. In addition, the cause candidate was upheld without judging/rejecting it, and multiple candidates were verified simultaneously. By referring to the behavioral model of the MEs who succeeded in troubleshooting, the following three patterns regarding the order of verification of the cause candidates were observed.

- Verification from the cause candidates that can be easily verified.
- Verification from the cause candidates related to the central part of the machine (which can cause troubles in a plurality of places).
- Verification from highly probable cause candidates.

Thus, for troubleshooting, it is important to know the cause candidates with high occurrence frequency and severity and verify the cause candidates in order of these viewpoints.

Method for Verifying Cause Candidates. For verifying whether the cause candidate is a true cause, the MEs who failed in troubleshooting only observed or were confused by the disassembling of all machine parts considered to be related. As a result, although the generated cause candidate was correct, some MEs misdiagnosed it in the verification process.

In contrast, the following patterns were observed in the verification method of successful MEs.

- They selectively used observation, palpation, and a simple test (e.g., passing air through a syringe to a filter) according to the cause candidate (They knew the verification method suitable for the cause candidate).
- They verified cause candidates only with the help of acquired information and their internal knowledge (know the clear verification criteria).

As information about the verification method, that is, the verification standard and verification order are not shown in the machine, it would be better to present the information based on these viewpoints.

4.3 Summary

Based on the above discussion, we modeled the problems in the interface design of dialysis machine and MEs, who are users of the machine (Fig. 7). It is necessary to study the interface design and analyze the education of MEs to improve these problems.

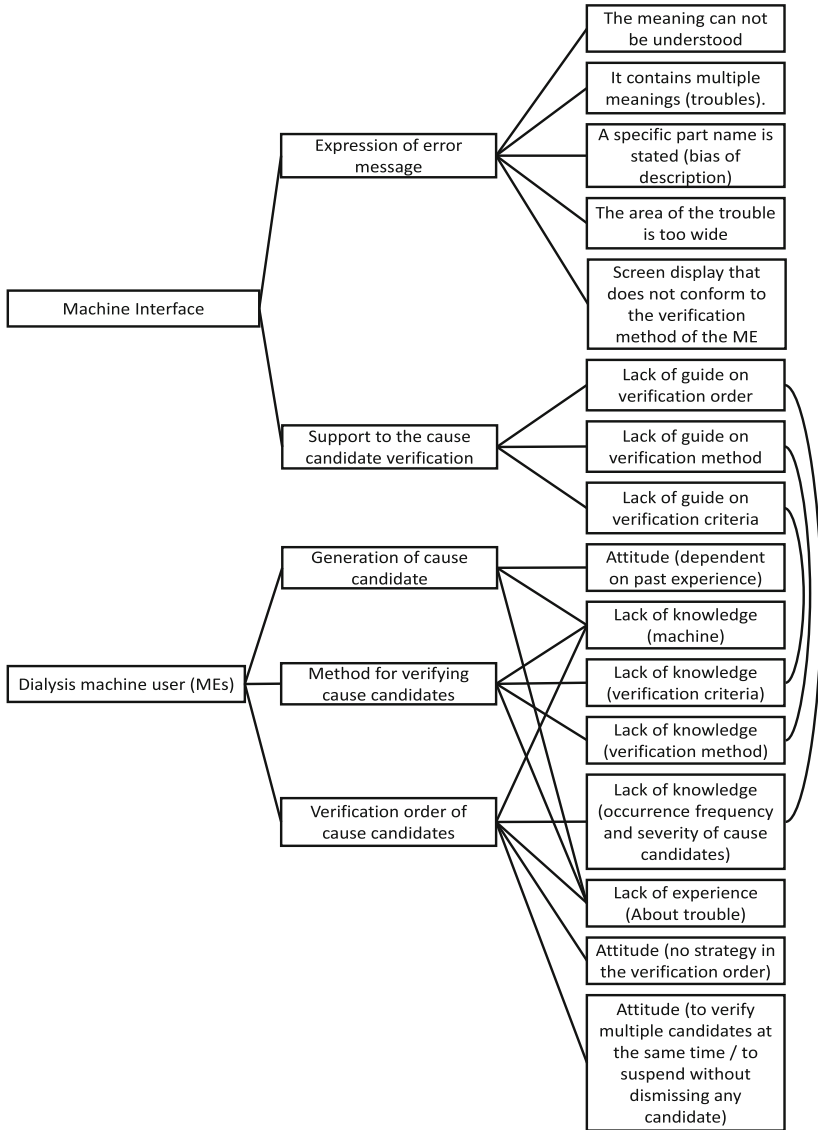


Fig. 7. Model of problems in interface design of dialysis machine and MEs.

5 Conclusion

With the aim of designing an interface that supports machine troubleshooting, an ME cognitive task analysis was conducted in this study. In addition, the error-message mechanism diagram and ME behavioral models regarding troubleshooting were clarified. We then compared the models, and clarified problems in the design interface for an alarm system of the machine and for MEs who are users of the machine, and then compiled these problems into a model. It is necessary to conduct a hearing survey of the MEs based on these diagrams, and study the type of alarm interface that should be designed using the obtained model.

By evaluating and improving the interface design related to the troubleshooting of the equipment by using the method of this research, it is expected that alarms for required contents will be displayed on the device when the user needs it.

References

1. Masakane, I., et al.: Annual dialysis data report 2015, JSDT registry. *Nihon Toseki Igakkai Zasshi* **50**(1), 1–62 (2017). doi:[10.4009/jstdt.50.1](https://doi.org/10.4009/jstdt.50.1)
2. Rinsyou kougaku goudou iinkai: Rinsyou kougaku gishi gyomubetsu gyomu shishin (Clinical engineers' treatment guidelines). Japan Association for Clinical Engineers, Tokyo (2012). (in Japanese)
3. Iryou kiki center: Iryou kiki no kiso chishiki (Basic knowledge of medical equipment), 2 edn. Yakuji Nippo, Tokyo (2008). (in Japanese)
4. Engineering Equipment & Material Users' Association (EEMUA): ALARM SYSTEMS A guide to Design, Management and Procurement, 2nd edn. EEMUA, London (2007). EEMUA Publication No. 191
5. Shneiderman, B., et al.: *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 6th edn. Pearson Education, London (2016)
6. Asakura, R., Katsumata, D., Tamaki, K.: Work instruction method based on the expected value minimization of failure cause identification working hours. *J. Jpn Ind. Manag. Assoc.* **67**(1), 37–48 (2016). doi:[10.11221/jima.67.37](https://doi.org/10.11221/jima.67.37)

Design of a Decision-Making Task for a Collaborative Brain-Computer Interface System Based on Emotiv EEG

Ânderson Schuh^(✉) and Márcia de Borba Campos

Faculty of Informatics (FACIN),
Pontifical Catholic University of Rio Grande do Sul (PUCRS),
Porto Alegre, Brazil

anderson.schuh@acad.pucrs.br, marcia.campos@pucrs.br

Abstract. This article presents lessons learned in the design, implementation and evaluation of a task of computerized decision-making to be used in a non-invasive collaborative and hybrid brain-computer interface, which used the Emotiv EEG for extract neural feature and response time as behavioral feature. The task developed was based on RSVP and has controlled levels of difficulty that can cause uncertainty. It is believed that the participants' general satisfaction was good, since the majority indicated that they had an easy understanding of the task. The task proved to be efficient for the initial purpose, that is, to generate difficulty to the participants and the experiment can be balanced with respect to the difficulty of executing the task. However, it was not possible to find relationships between the emotions felt by the participants in their subjective answers and in their emotions collected through the Emotiv EEG. It was possible to verify that the participants with less response time tend to answer more correctly, which can indicate their level of confidence, as expected.

Keywords: Collaborative BCI · Hybrid BCI · EEG · RSVP · Making-Decision · Response time

1 Introduction

A Brain-Computer Interface (BCI) measures directly the brain activity associated with the user's intention, and translates it into control signals which are detected and interpreted by applications [1–3]. The non-invasive BCIs can be based on the electroencephalogram signs (EEG), a device that distributes electrodes by the scalp and, through them, registers the electrophysiological brain activity [4]. A hybrid BCI (hBCI) consists of the combination of two or more types of BCI, two or more signal acquisition techniques, or a combination of a BCI with other interaction techniques not based on BCI [5–8]. But a Collaborative BCI (cBCI) integrates the brain activity of a group of individuals, with the main aim to improve the classification of signs or increase human capacity [9]. Similar to a conventional BCI, a cBCI system has key parts for its operation, and will be applied to two of them: signal acquisition and signal processing. First, the brain signals to a group of users are acquired by various recording devices and, then, are synchronized with environmental events which are common among all

the users. After this, the processing of the collected data takes place in order to decode the intentions of the users, and, finally, they are translated to operating commands [9].

Currently, BCI equipments are easily found on the market, as the case of Emotiv EEG and Neurosky MindWave. In spite of this, it only [10] used resources which are present in this equipment for the prototyping of a cBCI for humorous content classification in images. In this study, we used the Emotiv EEG.

A concern/limitation of the BCI area is the application of these systems in a real environment, especially with regard to the extension of human capacity, because other systems usually present more satisfactory results than a BCI [9]. In this way, it is necessary to find daily tasks that can benefit from its use. The Decision Making (DM) is an issue that has been discussed in the area of Economics and Administration [11]. In addition to being a day-to-day task, it shows, in many cases, complex problems, especially when there is risk or uncertainty [12]. A way to minimize those cases is the DM. However, some research shows that communication is not always an ally of this process [13]. With regard to the cBCIs, where the application domain is the decision making, there are studies that use the techniques of Go/NoGo [15, 16] and Rapid Serial Visual Presentation (RSVP) [17–19]. Through an RSVP, one can simulate adverse conditions to the user. Thus, to decrease the display time of an image at a rate where it is not possible to complete the identification of the elements, there is an uncertainty and the user needs to make a decision.

This article presents lessons learned in the design, implementation and evaluation of a task of computerized decision-making to be used in a non-invasive collaborative and hybrid brain-computer interface, which used the Emotiv EEG.

2 Method

We decided to perform an experiment similar to the one presented in [17], which developed a cBCI and used tasks based on RSVP. Despite the record in the RSVP design considerations, it was not considered the information coming from the intentional blink, saccadic movements and memorization after 6 s, which would require a longer time on the data analysis. Some important requirements were raised:

- After performing the task, ask the user for an auto report of his/her experience during the experiment, for possible triangulation of information with the neural and behavioral data collected.
- Some tasks are more emotional and can stimulate various mental processes. With this, the development of a simple task that doesn't involve emotional issues may facilitate the mapping of which processes are important.
- Use simple motor tasks, because, just as the stimuli, can interfere with the mapping that is important.
- Binary responses tasks are the most common and fit well in the original purpose of the work, as visual recognition tasks (search for targets) or planning.

2.1 Sample

Unlike what happens in the traditional research, the choice of the participants was intentional (not probabilistic), that is, they were chosen on the basis of the interest issues to the study, profile and availability to perform the tasks.

2.2 Data Collection and Instruments

The planning process of the data collection was based on five questions, suggested by [20]. It was analyzed the following characteristics of each participant of the research:

1. Profile by means of a questionnaire;
2. Neural Feature (NF) of the Affective Suite registered by the Emotiv EEG;
3. The Decision Making (DM) registered through the application containing the tasks;
4. Response Time (RT) registered through the application containing the tasks;
5. Perceptions by means of questionnaires.

The questionnaires were made available in digital form and were answered in the presence of the researcher, at specific times, following a given order. The log with the data collected by the BCI was automatically registered in a database.

It was used the Emotiv EEG and its Affectiv Suite for the acquisition of the NF, which monitors in real time the subjective emotions experienced by the user, based on brain waves, being possible to detect: engagement, meditation, frustration, instant excitement and long-term excitement. The detection of the engagement can investigate measures, such as surveillance, alertness, concentration, stimulation and interest. Instant Excitement detection can determine measures, such as excitement, nervousness and restlessness. Finally, the long-term excitement detection performs the same measurements as the previous detection, but its diagnosis is usually more accurate, since it analyzes these measures in a longer period of time (in minutes). The meditation detection represents the relaxation or stress level. The frustration measurement already recognizes what the own name says [21]. For all the detections, the values are represented in the scale of 0 to 1, considering 1 as a strong existence of the emotion, and 0 as the absence. Still, the same detection can point out different emotions, as the meditation detection, in which values close to 1 indicate mental relaxation, and around 0 represents stress or discomfort.

The task was developed using the game engine Unity 3D. For the storage of the collected data, it was used the embedded database SQLite. For the access to the Emotiv EEG Affectiv Suite data, it was used the software Mind your OSCs, which accesses the Emotiv engine, collects the data and makes them available through the OSC Protocol.

Each participant performed a sequence of 2 stages, each one containing 56 tests. Each test begins with the presentation of a white fixation cross centered on the computer screen with a duration of 1 s. This allows the participant to prepare him/herself for the presentation of the stimuli. After this, two screens are presented, each containing a set of images that are geometric shapes. The first set of images appears for about 80 ms and is immediately followed by a mask for 250 ms. The mask is used to erase any remnant of the first set. After 1 s, the second set for 100 ms is shown on.

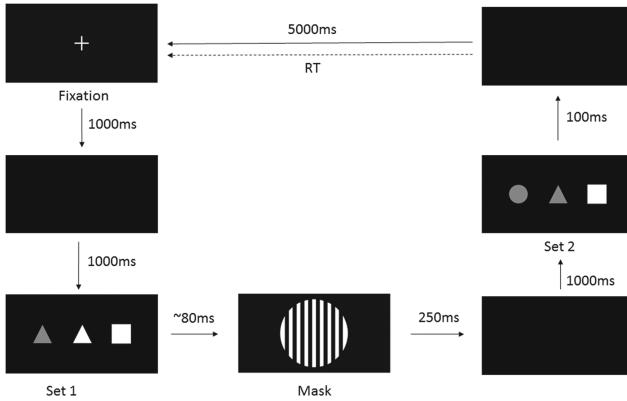


Fig. 1. Developed task RSVP based

As a result, the participant must decide, as soon as possible, if the two sets are identical or different. The answers are given through a conventional QWERTY standard keyboard positioned in front of the user. The F key must be used to indicate the same sets of images while the key J must indicate different sets. These keys were used because they have embossed marking, differentiating them in the keyboard. The response time of each decision is stored. The next test takes 5 s to begin, after the appearance of the second set of elements. Figure 1 illustrates this dynamic.

Each set consists of 3 geometric shapes, which can be any combination of triangles, squares and circles. Each shape can take one of the colors, white (1, 1, 1) or gray (0.65, 0.65, 0.65). With this, it can be said that each of the three elements in a set has two features, color and shape. Therefore, each set has a total of $(2 \times 3)^3 = 216$ different combination possibilities. If the two sets of one test are considered, we have $216^2 = 46.656$ possible combinations. In this sense, features may be shared, that is, when they occur in the same position in two sets of an essay.

In this way, each pair of a test set is sorted by the number of these shared features, which is called the Degree of Match (DoM). If all the elements of the first set are distinct from the elements of the second set, we have a $DoM = 0$. In the case of an element share a features, for example, the same color, we have a $DoM = 1$. In this way, the DoM of each test can range from 0 to 6. To exemplify, in Fig. 1, which has in the first position of the first set a gray triangle. In the second set, the first position has a gray circle. In this case, the two elements share a features, the color. Still, in the second set, the second element is a gray triangle. Although they present the same shape and color, the triangle of the first set does not share features with that one of the second set, since they are not in the same position.

The combination of elements of the first set is randomly generated. In order to produce a proportional experiment at the level of difficulty for the second set, some restrictions were adopted in the randomization of the combinations, using the DoM values to define the difficulty of each test, and, thus, to generate equal proportions of each DoM level in the experiment. In this case, the experiment had in its totality of tests a multiple value of 7, which represents the amount of possible DoM, totaling 96 different

sets and 16 identical sets. After the generation of the elements sets for the assays, they were stored to be used with all the participants, in order to perform the same experiment, increasing its repeatability and reproducibility. No set of stimuli is repeated.

The display time of the first set of elements stays below of what would normally be required for its complete visualization and perception, according to [22]. This can help to identify the confidence level of the participant, who, by failing to identify all the components of the set with high difficulty, may be in doubt, and present a longer response time, as described in the study of [17]. In comparison with [17], the waiting time after the response was increased, so that it was possible to register the features, mainly, of frustration and stress.

2.3 Procedures

The experiment was performed in a controlled environment, in a closed room with a table, in which were placed a portable computer that runs the applications, a 19-inch monitor, a mouse and a keyboard, an Emotiv EEG that was connected to the computer through the Wireless adapter, and two chairs. During the experiments, only the participant and the evaluator remained in the room. The laptop screen was used by the evaluator for monitoring, while the 19-inch monitor was used by the participant in the use of the task. All the evaluations were done individually.

For each of the evaluations, the previous configuration of the environment lasted about 40 min. This process took into account the organization of the environment, configuration of the applications and the preparation of the Emotiv. This stage is fundamental to the experiment because the bad hydration and the improper positioning of the electrodes can generate noise in the signals captured by the Affective Suite, or the lack of the Emotiv communication with the application of the data collection or the application of the tasks can generate the lack of some data registration.

With the configured environment, the researcher explained to each participant the research objectives, data collection procedures, the collected data confidentiality, estimated duration of the experiment, discomforts that might be felt during the tasks, among other informations. The participation only occurred after the participant's agreement manifestation, in the informed consent of the research.

The pre-test questionnaire was applied and, afterwards, the researcher positioned the Emotiv on the participant's head, then he tested if all the electrodes were correctly placed and explained the task that was to be performed. The equipment was placed before the actual task was done, so that during the explanation of the task, it was self-calibrated to the participant, as recommended by the equipment manufacturer. Then, the participant positioned himself in front of the monitor, with the index fingers on the keys F and J. After that, the evaluator started the task. After performing the first 56 attempts, the user answered the questionnaire regarding his/her perception about the first stage. After completion, the volunteer performed the second stage of the experiment and answered the second questionnaire. Finally, the participant completed the final questionnaire about his/her overall perception of the experiment. Only at the end is that the equipment was removed from the participant's head. This process took about 40 min for each participant.

3 Results

Initially, the sample was composed of 11 participants who answered the questionnaires and performed the activities. However, in the analysis of the database, it can be seen that all the Emotiv EEG data coming from a participant had not been recorded. Therefore, this participant was disregarded of the final sample. On the other hand, we chose to keep the data of another participant who did not only have the recorded frustration data. Thus, the final sample consisted of 10 participants.

3.1 Profile of the Participants

The questionnaire regarding the participants' profile contained 8 mandatory questions, besides the name and an identifier, which were filled out by the researcher.

Of the 10 participants of the experiment, 9 were male and 1 was a female. 8 participants are between 18 and 27 years and 2 participants are between 28 and 37 years. As for nationality, 9 were Brazilians and one was Colombian. None of the participants indicated to be photosensitive, thus, do not have sensitivity to light.

As for the use of the hands, all of them indicated to be right-handed and informed that the right hand is the one with which, preferably, they handle the mouse and the mouse buttons. Regarding the general use of the keyboard, 4 participants reported using both hands and all their fingers and write without looking at the keyboard, 2 reported using their left hand and all their fingers, being that one of them does not look to the keyboard while typing, 2 participants use only their right hand and all their fingers, and 2 others did not report the hand that they usually use to write, but reported typing without looking at the keyboard. To write messages on the smartphone, 8 participants indicated that they use both hands and, predominantly, a finger of each hand, and 2 participants indicated that they use only they right hand and, predominantly, a finger of that hand to write the messages. When asked about the frequency of the perception of games use in the week before the experiment, the majority (60%) reported that they did not use this type of game, 20% played 3-4 days, 10% played every day and 10% played for 1-2 days.

3.2 Questionnaire Stage 1

The explanation of the questions in the Questionnaire Stage 1 was made prior to the activities of the Stage 1. This procedure was adopted in order that the participant had knowledge of what he/she should answer and observe during this stage. At all times, the participant remained with the BCI connected and in operation to prevent it from being recalibrated.

This questionnaire had 6 questions, 4 of which were of multiple choice, 1 with answers based on semantic differential scales, and the other one was open, which are discussed below.

Have you felt any visual discomfort? (Fig. 2): participant 3 reported that he had visual discomfort but, even though he did not specify the timing, it is believed to have

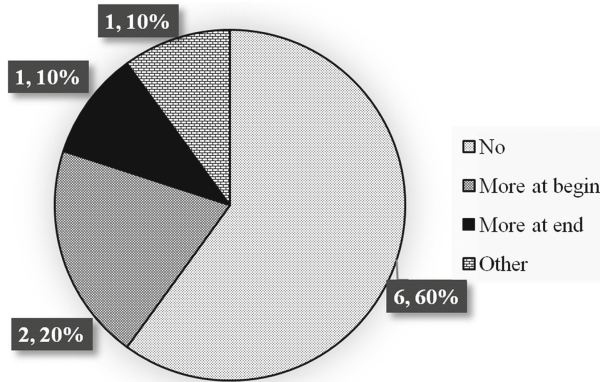


Fig. 2. Have you felt any visual discomfort?

been during the course of the task because he suggested that the ambient light could be darker. Moreover, comparing the participants' responses with their accuracy and error rates between the stages of the experiment, it cannot be said that such sensations perceived by the participants reflected in their performance.

In general, about the use of keys in the keyboard to indicate the choices "same" and "different", do you consider that it was: the range of answers had a scale of 1 to 5, representing from the more difficult (1) to the easier one (5). The average of the answers was 4.1, indicating that the use of the F and J keys were found to be easy to use.

Participant 5 was the only one who considered the use of the keys more difficult than easy, even though he/she indicated that he/she makes use of the keyboard without looking at the keys. It is noteworthy that this participant was the one that obtained the best rate of success in Stage 1, with 91% of success (51 questions). He/she was the participant who obtained the best overall score (88.4%–99 questions). The participants who considered the F keys and J easy to be used had their levels of accuracy rates near average, just above or below.

Overall, how did you feel during the challenges? This question was divided in 8 questions, applied with semantic differential scales, as follows, and the answers from 1 to 5:

- Distracted (1) × Tuned (5): the average was 4.1, demonstrating that the participants felt themselves to be more attentive than distracted on the achievement of the challenges. Only the Participant 5, who considered the use of the keys harder to be used, indicated that he/she was more distracted than attentive in performing Stage 1;
- Uncompromised (1) × Engaged (5): the average was 4.4, demonstrating that the participants felt themselves to be more engaged than uncompromised while performing the challenges;
- Frustrated (1) × Satisfied (5): considering that 3 is the central point of the scale and the average was 3.4, it demonstrates that the participants felt themselves a little happier than uncompromised while performing the challenges. Only Participant 2

indicated that he/she was frustrated while performing the activities, however, his/her accuracy rate in Stage 1 remained on the average (82.1% and 46 issues);

- Bored (1) × Excited (5): the average was 4.2, demonstrating that the participants felt themselves more excited than bored while performing the challenges;
- Stressed out (1) × Relaxed (5): considering the central point of the scale and that the average was 3.3, it demonstrates that the participants felt themselves a little more relaxed than stressed while performing the challenges. Participants 2, 5, 11 and 12 indicated that they were more stressed than relaxed while performing Stage 1. Of these, Participant 12 presented the lowest accuracy rate, being of 58.9%, which is equivalent of 33 hits. At the end of the experiment, its average rose to 67.9%, with a total of 76 hits;
- Nervous (1) × Calm (5): the average was 3.7, demonstrating that the participants felt themselves a little calmer than nervous while performing the challenges. Participants 2, 5 and 11 indicated that they were more nervous than calm, and had already indicated they were more stressed than relaxed;
- Angry (1) × Quiet (5): the average was 4.2, demonstrating that the participants felt themselves more calm than angry while performing the challenges. Only Participant 2 indicated to be angrier than calm;
- Uncertain (1) × Confident (5): the average was 3.8, demonstrating that the participants felt themselves confident than uncertain while performing the challenges. Participants 2 and 5 indicated that they were uncertain than confident while performing the activities of Stage 1. On the other hand, if the accuracy rate of the Participant 2 was 82.1% 2 and 46 hits, which was the average of the participants in Stage 1. Participant 1 was the one that got the best accuracy rate in Stage 1, which was of 91.1% and 51 hits.

Overall, do you think that your accuracy rate was better at identifying the “same” or “different”? As previously described, this stage had 56 pairs of screens (stimulus set), 8 of which represented the same screens and 48 different screens. Most participants (80%) (Participants 2, 3, 4, 5, 6, 7, 9, 10, 11) consider that their accuracy rate was better in identifying the different, one participant (10%) (Participant 12) considers that his/her rate was better in identifying the same and the other (10%) (Participant 6) considers that there is no significant difference in identifying the same or the different. These results demonstrate that the participants are aware of the difficulty to observe small details in stimuli, demonstrating the efficiency of the proposed task.

Do you think that your accuracy rate was better in that phase of Stage 1? The majority (80%) (Participants 2, 3, 4, 6, 9, 10, 11 and 12) consider that their accuracy rate was better in the final phase of the stage, in keeping with the already informed statements. Of these, only Participant 3 did not have his/her answer confirmed, being that in the first half of Stage 1, he/she made 23 hits, and in the second half he/she hit 21 times. One (10%) (Participant 5) found that there was no significant difference in his/her accuracy rate, confirmed by the correct answers, being that 25 were in each half of Stage 1. Another (10%) (Participant 7) could not state, being that there was an increase of 6 hits in the final half of Stage 1.

3.3 Questionnaire Stage 2

The procedure adopted in the second stage was the same from the previous stage, that is, before the battery of tests was started, the questionnaire was read and explained, which contained 6 identical questions to the previous questionnaire, which are discussed below.

Have you felt any visual discomfort?

Participants were asked if they felt uncomfortable sensations during the visual task. The answers are illustrated in Fig. 3.

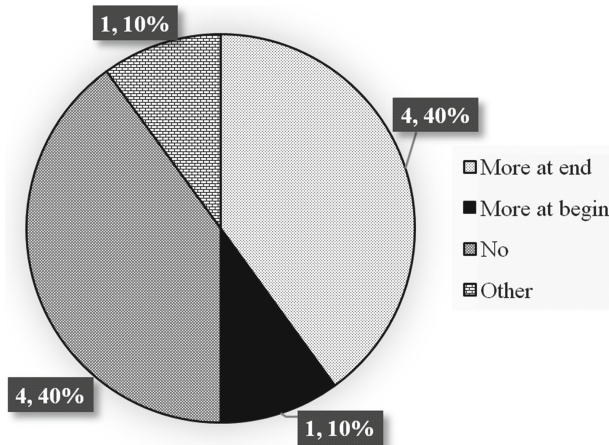


Fig. 3. Have you felt any visual discomfort?

4 participants (40%) (Participants 2, 4, 6 and 7) reported that they had more visual nuisance at the end of the stage. Of these, Participant 2 had also indicated that he/she had it only at the end of Stage 1, Participants 4 and 7 indicated that they had no nuisance in Stage 1, and Participant 6 had indicated that he/she had at the beginning of Stage 1.

Four participants (40%) (Participants 5, 9, 10 and 11) had no nuisance. Of these, Participants 5, 10 and 11 also did not indicate nuisances in Stage 1. And Participant 9 had indicated nuisance at the beginning of Stage 1.

Participant 1 (10%) (Participant 12) reported that he/she had nuisance at the beginning of the stage, being that he/she indicated no nuisance during Stage 1.

The other one (10%) (Participant 3) did not specify in what time he/she had visual nuisance, but continued recording suggestion so that the test environment was less bright.

Overall, about the use of keys on the keyboard to display the choices “same” and “different”, do you consider that it was: The answers range had a scale of 1 to 5, representing from the more difficult (1) to the easier one (5). The answers average was of 4.4, being that most of the participants kept their assessments. Participants 5 and 12 considered that the use of these keys got easier in Stage 2. Thus, considering that the average of Stage 1 was of 4.1, it can be inferred that the use of the keys throughout the experiment facilitated their learning and use.

Overall, how did you feel during the challenges? This question was divided in 8 questions, applied with semantic differential scales, as follows, and the answers were registered from of 1 to 5:

Distracted (1) \times Tuned (5): the average was 4.3, demonstrating that the participants felt themselves to be more attentive than distracted on the achievement of the challenges. Only the Participants 4 and 6 indicated having diminished their attention, from 4 to 3 and from 5 to 3, respectively. The other ones indicated or an increase (Participants 3, 5, 9 and 11) or kept the same level from Stage 1 (Participants 2, 7, 10, 12);

- Uncompromised (1) \times Engaged (5): the average was 4.6, demonstrating that the participants felt themselves to be more engaged than uncompromised while performing the challenges;
- Frustrated (1) \times Satisfied (5): considering that 3 is the central point of the scale and the average was 3.7, it demonstrates that the participants felt themselves a little happier than uncompromised while performing the challenges;
- Bored (1) \times Excited (5): the average was 4.2, demonstrating that the participants felt themselves more excited than bored while performing the challenges;
- Stressed out (1) \times Relaxed (5): considering the central point of the scale and that the average was 3.6, it demonstrates that the participants felt themselves a little more relaxed than stressed while performing the challenges;
- Nervous (1) \times Calm (5): the average was 3.6, demonstrating that the participants felt themselves a little calmer than nervous while performing the challenges;
- Angry (1) \times Quiet (5): the average was 4, demonstrating that the participants felt themselves more calm than angry while performing the challenges;
- Uncertain (1) \times Confident (5): the average was 4, demonstrating that the participants felt themselves confident than uncertain while performing the challenges.

Overall, do you think your hit rate was better in identifying the "same" or "different"? Most of the participants (80%) (Participants 2, 4, 5, 7, 9, 10, 11 and 12) consider that its accuracy rate was better in identifying the different, one participant (10%) (Participant 3) considers that its rate was better in identifying the same and other participant (10%) (Participant 6) considers that there is no significant difference in identifying the same or the different. As in Stage 1, the participants demonstrate awareness of the difficulty to observe small details in stimuli, demonstrating the efficiency of the proposed task.

Do you think that your accuracy rate was better in which phase of the Stage 1? Three participants (30%) (Participants 9, 11 and 12) believe that their accuracy rate was better in the early phase of the stage, however, they had in average 4.6 hits in the final half of Stage 2. Three participants (30%) (Participants 4, 6 and 10) believe that there was no significant difference in their accuracy rate, in spite of this, the participants had an increase at the end of Stages 5 and 8 and 1 hit, respectively. One participant (10%) (Participant 3) thinks that he/she was better between the phases, which was validated, and between the Test 70 and the 88, he/she obtained 25 hits, against 23 hits in the other tests of Stage 2. One participant (10%) (Participant 2) believes that his/her accuracy rate was better in the final phase and other participant (10%) (Participant 7) didn't know what to inform.

3.4 Final Questionnaire

Upon completion of Stage 2, and completing the questionnaire relating to that stage, a final questionnaire was applied.

Do you consider that the more challenges similar to those that have been presented have to solve, the better your attention to observe visual sequences? Eight participants (80%) (Participants 2, 3, 5, 7, 9, 10, 11 and 12) consider that the greater the experience with the task, the higher their attention. One participant (10%) (Participant 4) believes that there would be no improvement in attention and other one (10%) (Participant 6) believed to be indifferent. However, one can't see this performance improvement during the experiment.

Do you consider that the more similar challenges to those which were presented you have to solve, the better will be your agility to use the keys of the keyboard to indicate "same" and "different"? Six participants (60%) (Participants 2, 4, 5, 7, 9, and 12) consider that the greater the experience with the task, the higher will be their agility with the keys. Three participants (30%) (Participants 3, 6 and 10) positioned themselves as indifferent to the improvement. Finally, one participant (10%) (Participant 11) considers that it wouldn't make any difference. It was found that there is a small improvement in the average answer time of the participants between the stages, being 923 ms in stage 1 and 870 ms in stage 2.

Considering that it is a visual attention challenge, do you consider that the time that separates one challenge from another (marked by the screen +) was: Eight participants (80%) (Participants 2, 3, 4, 5, 6, 7, 9 and 12) consider that time is suitable and 2 participants (20%) (Participants 10 and 11) indicated that it could be higher.

Considering that it is a challenge of visual attention, do you consider that the time to identify the sequences same and different was: The majority of the participants (70%) (Participants 2, 3, 4, 5, 6, 10 and 12) consider that the time could be higher, while 3 participants (30%) (Participants 7, 9 and 11) believe that the time was appropriate. This confirms that the participant's perception goes against the difficulty proposed in the task.

Considering that it is a challenge of visual attention, do you consider that the time to identify the sequences same and different should vary along the Stage? For example, start with a longer duration and reduce it until it stabilizes. Nine participants (90%) (Participants 2, 3, 4, 5, 6, 9, 10, 11 and 12) considered that time could be reduced according to the experience with the task, but 1 participant (10%) (Participant 7) felt that it could not be reduced.

Considering that it is a challenge of visual attention, do you consider that the background colors and objects are clearly identified? Five participants (50%) (Participants 6, 7, 9, 11 and 12) consider that the colors used are clearly identified. Four participants (40%) (Participants 2, 3, 5 and 10) consider that the colors are not easily identified and, finally, one participant (10%) (Participant 4) considers it as indifferent. The difficulty in identifying the colors is related to the proposed time of exhibition, however, the majority of the participants were able to notice these differences.

3.5 Relationship Between Measures and Responses

In the previous sections, we sought to present the participants' perception to perform the experiment, data resulting from the data collection by means of a questionnaire. This section seeks to present the data collected through the BCI and discuss possible relations between neural and behavioral measures, as well as the participants' perceptions. It also includes discussion of error rates and relation to the difficulties of the proposed task.

As for the data collected regarding to the neural measurements, each test generated, on average, 70 records of each of the 5 measures extracted by the Emotiv EEG, which are engagement, frustration, meditation, excitement, and long-term excitement. Also, the response times of each test were stored.

Due to the generated data volume, it was performed summarization of the neural feature data per assay. The summarization is a descriptive statistics area technique, which consists of the synthesis of the data collected and, despite the loss of informations, it is still a minor factor compared to the gain provided in the interpretations [23]. The average will be used as a summarization method, since it is tried to understand if the increase or decrease of some of the neural features can influence in the decision making.

In Table 1 are presented the average of each of the neural and behavioral measures, along with its standard deviation, minimum and maximum value. Due to a read error, the frustration data of Participant 7 were not recorded correctly and not to interfere with the general analysis, the frustration data of this participant were not considered in the analysis. Therefore, the other features are based on data of 10 participants while the frustration is based on the data of 9 participants. It turns out that, considering the central measures and the standard deviation, the analyzed periods do not have considerable differences.

Table 1. General information about the recorded measures, * 9 data users.

Feature	Average	Median	STD	Minimum	Maximum
Engagement	0.5712	0.5750	0.1187	0.0984	1.0000
Excitement	0.3360	0.3010	0.1552	0.0214	0.9885
Meditation	0.3811	0.3659	0.0636	0.2149	0.6667
Frustration*	0.4599	0.4177	0.1734	0.1203	1.0000
Response time	0.8967	0.8000	0.4155	0.0800	3.8600

Table 2 summarizes the amounts of errors and hits by difficulty level of all users. There are 112 trials per participant, distributed equally in the DoM 7 levels, from 0 to 6. It is verified that the greater the DoM, the greater the difficulty of the user to perceive the details of the image, and, in this way, greater the error rate. On the other hand, it is verified that DoM 6 has more hits than DoM 5. This occurs, since the user, when he/she cannot identify a different features with a higher incidence in DoMs 5 and 6, ended up voting as if it were the same. In the case of DoM 6, the answer is considered correct. In this way, it is believed that the applied theory of RSVP in the proposed task, fulfills its role, generating the difficulty in recognizing the more complex images.

Table 2. Accuracy rate and error by levels of DoM

Difficulty	Accuracy	Errors	Not rated
0	156	4	0
1	150	10	0
2	153	6	1
3	131	28	1
4	126	34	0
5	89	71	0
6	109	51	0

Table 3 presents the accuracy and error rates of the participants in stages 1 and 2. The data has been leased to a decimal place. It can be verified that 2 participants (2 and 9) had a small increase in the accuracy rate (both from 82.1% to 83.9%), being only one question. Three participants (7, 11 and 12) had greater increases in the accuracy rate (from 78.6% to 87.5%, from 73.2% to 85.5%, and from 58.9% to 76.8%, respectively), with 5, 7 and 10 questions, respectively. Three participants (3, 4, 10) had a slight increase in the error rate (from 21.4% to 23.2%, from 14.3% to 16.1% and from 16.1% to 17.9%, respectively), being only of one question, while one participant (5) had an increase of 3 wrong questions (from 8.9% to 14.3%). One participant (6) maintained his/her score in two stages, with 85.7% of accuracy.

Table 3. Accuracy rate per participant between stages

	Stage 1				Stage 2				Total			
	Accuracy		Error		Accuracy		Error		Accuracy		Error	
ID	Qty.	%	Qty.	%	Qty.	%	Qty.	%	Qty.	%	Qty.	%
2	46	82.1	10	17.9	47	83.9	9	16.1	93	83.0	19	17.0
3	44	78.6	12	21.4	43	76.8	13	23.2	87	77.7	25	22.3
4	48	85.7	8	14.3	47	83.9	9	16.1	95	84.8	17	15.2
5	51	91.1	5	8.9	48	85.7	8	14.3	99	88.4	13	11.6
6	48	85.7	8	14.3	48	85.7	8	14.3	96	85.7	16	14.3
7	44	78.6	12	21.4	49	87.5	7	12.5	93	83.0	19	17.0
9	46	82.1	10	17.9	47	83.9	9	16.1	93	83.0	19	17.0
10	47	83.9	9	16.1	46	82.1	10	17.9	93	83.0	19	17.0
11	41	73.2	14	25.0	48	85.7	8	14.3	89	79.5	22	19.6
12	33	58.9	22	39.3	43	76.8	13	23.2	76	67.9	35	31.3
avg	46	82.1	10	17.9	47	83.9	9	16.1	93	83.0	19	17.0
std	5.0	8.9	4.6	8.3	2.1	3.7	2.1	3.7	6.4	5.7	6.1	5.4

Still, the next questions addressed the user’s perception during the execution of the experiment. Table 4 present a summary of the data collected.

Table 4. Subjective answers per participant (ID) ans per Stage 1 (S1) and Stage 2 (S2).

ID	Dis. x Tun.		Unc. x Eng.		Frustr. x Sat.		Bor. x Exc.		Str. x Rel.		Ner. x Cal.		Ang. x Qui.		Uns. X Con.	
	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
2	4	4	5	5	2	2	4	5	2	2	2	2	2	2	2	2
3	4	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5
4	4	3	5	5	4	5	5	5	5	5	5	5	5	5	4	5
5	2	4	3	4	3	3	4	3	2	2	2	2	3	3	2	4
6	5	3	3	4	3	4	4	4	4	4	3	4	5	4	5	4
7	4	4	5	5	3	4	4	4	3	5	5	4	4	4	4	4
9	4	5	4	4	3	3	4	4	4	3	4	3	4	3	4	4
10	5	5	5	5	3	4	5	5	4	4	4	5	5	5	4	4
11	4	5	5	5	5	5	4	4	2	2	2	2	5	5	5	5
12	5	5	4	4	3	3	3	3	2	4	5	4	4	4	3	3
avg	4.1	4.3	4.4	4.6	3.4	3.7	4.2	4.2	3.3	3.6	3.7	3.6	4.2	4	3.8	4

In order to compare the neurophysiological measures with the participants’ subjective answers, a table of the same scale of the collected neurophysiological measures was created. Initially, for each participant, the maximum and minimum values of each of the neural variables were found. After that, the maximum by the minimum was subtracted, and the result was divided by 5, in order to obtain the extreme values of each of the 5 levels of the scale. After that, the average of each of the neural variables was extracted for each of the stages. Afterwards, these values were classified in a scale of 1 to 5, based on the values obtained in the first operation. In the case of the neural measure Frustration, the values were inverted, being the lowest value 5 and the higher 1. This was necessary to adjust the scale used in the questionnaire that varied from Frustrated (1) to Satisfied (5). The obtained values are shown in Table 5.

General Engagement was compared to Distracted vs. Tuned and Uncompromised × Engaged, general Excitement was compared to Bored vs. Excited, general Meditation was compared to Stressed vs. Relaxed, Nervous vs. Calm and Irritated vs. Quiet, and lastly, general Frustration with Frustrated vs. Satisfied.

From 160 possible answers (8 subjective questions in 2 stages to the 10 participants), 4 participants had no equal value between the subjective and the neural answers. Only 15 answers of all participants obtained the same values between the subjective responses and the neural variables, however, these were not repeated between the stages. As an example, Participant 2, for the subjective answer of the Stage Distracted vs. Tuned, got level 4, the same level of his/her neural feature Engagement for Stage 1. When analyzing the increase or decrease in the levels between stages, we noticed 22 times of the 84 possible changes, being that from the 22 times, 12 times remained the

Table 5. General levels for neural features per participant (ID) ans per Stage 1 (S1) ans Stage 2 (S2).

ID	Gen. engagement		Gen. excitement		Gen. meditation		Gen. frustration	
	S1	S2	S1	S2	S1	S2	S1	S2
2	4	3	2	2	1	2	1	2
3	3	2	2	2	3	3	3	1
4	3	2	2	2	3	3	2	3
5	2	3	3	2	1	2	3	2
6	3	2	3	3	2	2	3	3
7	4	2	1	2	2	4		
9	4	2	2	2	2	2	3	2
10	3	2	2	2	2	3	2	2
11	3	3	2	3	3	3	2	2
12	3	1	2	1	4	3	1	1
avg	3.20	2.20	2.10	2.10	2.30	2.70	2.22	2.00

same values for stage 1 and stage 2. As an example, participant 4, for the subjective answer Distracted vs. Tuned, had a variation of 1 level between stage 1 (level 4) and stage 2 (level 3), as to the levels of the neural feature Engagement, also occurred a decrease of 1 level between stage 1 (level 3) and stage 2 (level 2), i.e., the same behavior.

When analyzing the participants' average, a similar behavior between the subjective answers and the neurophysiological measures can be noticed in two cases. As for the General Excitement, 2.1 was obtained in both stages, being that in the "Bored \times Excited" questionnaire it was obtained 4.2 in both stages. Still, in the case of general Meditation regarding the question "Stressed \times Relaxed", the two participants had an increase between 0.3 and 0.4 points. However, for both cases, physiological measures presented medium-low levels and the subjective answers presented medium-high levels.

In this sense, we sought to analyze the relationship between the subjective answer for "Uncertain \times Confident" with the participant's accuracy rate. Participant 5 got the best accuracy rate, reaching 88.4%, however, it was not the participant who said to be confident. In the first stage, his/her accuracy rate was 91.1%, and his/her subjective answer was 2, the lowest recorded. The same participant also indicated an improvement in his/her security, raising from 2 to 4 in the second stage, however, his/her accuracy rate decreased to 85.7%. Participant 4 indicated an increase of his/her security from 4 to 5 between stage 1 and stage 2. On the other hand, his/her security rate fell from 85.7 to 83.9, i.e. a hit less. Participant 6 indicated a drop from 5 to 4 points in his/her security between the stages, but his/her accuracy rate remained at 85.7% in both stages. Participants 7, 11 and 12 obtained the highest accuracy rate variations between the both stages, however, their perceptions among the stages remained the same. In this way, no evidence was found that could satisfactorily relate the participants' subjective answers

that were collected through questionnaires with the neural measurements, which were recorded by the BCI through the adopted formulas.

The RT was analyzed using boxplot graphs (Fig. 4), where it was found that the values for correct answers are concentrated in lower values than in the wrong answers. This behavior was expected, as previously described, since participants with greater certainty make decisions more quickly. However, we can see that most of the values still share the same distribution. This may indicate that the participants are right even in doubt.

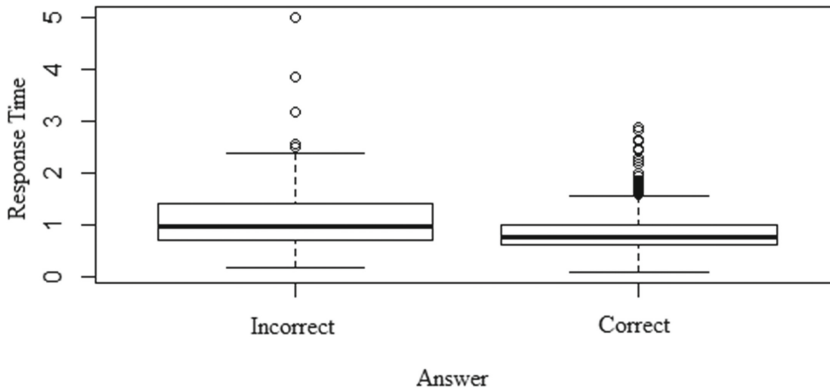


Fig. 4. Response times boxplot graph for all participants

4 Conclusion

This article presented lessons learned from the design, implementation and evaluation of a computerized decision-making task for use in a non-invasive, collaborative and hybrid brain-computer interface using Emotiv EEG. The BCI area has advanced in the last few years, especially in new approaches such as hybrid BCIs and collaborative BCIs. End-user equipment, such as Emotiv EEG, facilitates the access to the signal acquisition equipment.

It is believed that the participants' general satisfaction was good, since the majority indicated that they had an easy understanding of the task. As for the visualization time of the stimuli, the task proved to be efficient for the initial purpose, that is, to generate difficulty to the participants, along with the DoMs. In this way, the experiment can be balanced with respect to the difficulty of executing the task.

However, it was not possible to find relationships between the emotions felt by the participants in their subjective answers and in their emotions collected through the Emotiv EEG's Affective Suite. Moreover, it was possible to verify that, in an empirical way, the participants with less response time tend to answer more correctly, which can indicate their level of confidence, as expected.

References

1. Wolpaw, J.R.: Brain-computer interfaces as new brain output pathways. *J. Physiol. Online* (2007). doi:[10.1113/jphysiol.2006.125948](https://doi.org/10.1113/jphysiol.2006.125948)
2. Leuthardt, E.C., Schalk, G., Roland, J., Rouse, A., Moran, D.W.: *Neurosurgical Focus* (2009). doi:[10.3171/2009.4.FOCUS0979](https://doi.org/10.3171/2009.4.FOCUS0979)
3. Graimann, B., Allison, B., Pfurtscheller, G.: Brain-computer interfaces: a gentle introduction. *Brain-Comput. Interfaces* (2010). doi:[10.1007/978-3-642-02091-9_1](https://doi.org/10.1007/978-3-642-02091-9_1)
4. Lebedev, M.A., Nicolelis, M.A.: Brain-machine interfaces: past, present and future. *Trends Neurosci.* (2006). doi:[10.1016/j.tins.2006.07.004](https://doi.org/10.1016/j.tins.2006.07.004)
5. Wolpaw, J.R.: Brain-computer interfaces as new brain output pathways. *J. Physiol.* (2007). doi:[10.1113/jphysiol.2006.125948](https://doi.org/10.1113/jphysiol.2006.125948)
6. Nicolas-Alonso, L.F., Gomez-Gil, J.: Brain computer interfaces, a review. *Sensors* (2012). doi:[10.3390/s120201211](https://doi.org/10.3390/s120201211)
7. Amiri, S., Fazel-Rezai, R., Asadpour, V.: A review of hybrid brain-computer interface systems. *J. Adv. Hum.-Comput. Interact. Spec. Issue Using Brain Waves Control Comput. Mach.* (2013). doi:[10.1155/2013/187024](https://doi.org/10.1155/2013/187024)
8. Pfurtscheller, G., et al.: The Hybrid BCI. *Front. Neurosci.* (2010). doi:[10.3389/fnpro.2010.00003](https://doi.org/10.3389/fnpro.2010.00003)
9. Wang, Y., Wang, Y.-T., Jung, T.-P., Gao, X., Gao, S.: A collaborative brain-computer interface. In: *Proceedings of 2011 4th International Conference on Biomedical Engineering and Informatics* (2011). doi:[10.1109/BMEI.2011.6098286](https://doi.org/10.1109/BMEI.2011.6098286)
10. Stoica, A.: MultiMind: multi-brain signal fusion to exceed the power of a single brain. In: *Proceedings of the 3rd International Conference on Emerging Security Technologies* (2012). doi:[10.1109/EST.2012.47](https://doi.org/10.1109/EST.2012.47)
11. Chiavenato, I.: *Administração: Teoria, Processo e Prática*. Makron Books (2000). ISBN:85.346.1078-9
12. Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., Frith, C.D.: Optimally interacting minds. *Science* (2010). doi:[10.1126/science.1185718](https://doi.org/10.1126/science.1185718)
13. Manes, F., Sahakian, B., Clark, L., Rogers, R., Antoun, N., Aitken, M., Robbins, T.: Decision-making processes following damage to the prefrontal cortex. *Brain J. Neurol.* (2002). doi:[10.1093/brain/awf049](https://doi.org/10.1093/brain/awf049)
14. Hagen, G.F., Gatherwright, J.R., Lopez, B.A., Polich, J.: P3a from visual stimuli: task difficulty effects. *Int. J. Psychophysiol.* (2006). doi:[10.1016/j.ijpsycho.2005.08.003](https://doi.org/10.1016/j.ijpsycho.2005.08.003)
15. Valeriani, D., Poli, R., Cinel, C.: A collaborative brain-computer interface to improve human performance in a visual search task. In: *International IEEE/EMBS Conference on Neural Engineering* (2015). doi:[10.1109/NER.2015.7146599](https://doi.org/10.1109/NER.2015.7146599)
16. Yuan, P., Wang, Y., Gao, X., Jung, T.-P., Gao, S.: A collaborative brain-computer interface for accelerating human decision making. In: Stephanidis, C., Antona, M. (eds.) *UAHCI 2013. LNCS*, vol. 8009, pp. 672–681. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-39188-0_72](https://doi.org/10.1007/978-3-642-39188-0_72)
17. Poli, R., Valeriani, D., Cinel, C.: Collaborative brain-computer interface for aiding decision-making. *PLoS one* (2014). doi:[10.1371/journal.pone.0102693](https://doi.org/10.1371/journal.pone.0102693)
18. Poli, R., Cinel, C., Sepulveda, F., Stoica, A.: Improving decision-making based on visual perception via a collaborative brain-computer interface. In: *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA 2013* (2013). doi:[10.1109/CogSIMA.2013.6523816](https://doi.org/10.1109/CogSIMA.2013.6523816)

19. Valeriani, D., Poli, R., Cinel, C.: A collaborative brain-computer interface for improving group detection of visual targets in complex natural environments. *International IEEE/EMBS Conference on Neural Engineering* (2015). doi:[10.1109/NER.2015.7146551](https://doi.org/10.1109/NER.2015.7146551)
20. Rogers, Y., Sharp, H., Preece, J.: *Design de interação: além da interação humano-computador*, 3rd edn. Bookman, Porto Alegre (2013)
21. Emotiv, July 2016. www.emotiv.com
22. Spence, R., Witkowski, M.: *Rapid Serial Visual Presentation: Design for Cognition*. Springer, London (2013). doi:[10.1007/978-1-4471-5085-5](https://doi.org/10.1007/978-1-4471-5085-5)
23. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E.: *Multivariate Data Analysis*. Prentice Hall, New Jersey (2005)

Effects of Key Size, Gap and the Location of Key Characters on the Usability of Touchscreen Devices in Input Tasks

Da Tao¹(✉), Qiugu Chen¹, Juan Yuan¹, Shuang Liu²,
Xiaoyan Zhang¹, and Xingda Qu¹

¹ Institute of Human Factors and Ergonomics,
College of Mechatronics and Control Engineering, Shenzhen University,
Shenzhen, China

{taoda, zhangxyan, quxd}@szu.edu.cn,
chen_qiugu@163.com, 2515809640@qq.com

² Marine Human Factors Engineering Lab,
China Institute of Marine Technology and Economy, Beijing, China

Abstract. Touchscreen technology has gained increasing popularity over the last decade in a variety of personal, public and occupational settings. It is of great significance to investigate the effects of interface design factors that may affect the use of the technology. This study was conducted to investigate the effects of key size (ranged from 10 to 25 mm with 5-mm increments), gap (presence and absence) and the location of key characters (upper left, upper right, central, lower left and lower right) on usability metrics (i.e., task completion time, accuracy rate and user preference) in touchscreen input tasks. Fourteen undergraduate students (8 male and 6 female) participated in this study and were required to complete letter input tasks. The results indicated that there was a significant effect of key size on task completion time and accuracy rate, while gap and the location of key characters yielded no measurable effect on user performance. The performance was better for larger key sizes (≥ 15 mm) than smaller ones. The location of key characters significantly interacted with gap on accuracy rate. Users preferred 15 mm key size, the presence of gap and centrally located key characters. The results may help with the design of more usable and safe touchscreen technology.

Keywords: Touchscreen · Key size · Gap · Location of key characters

1 Introduction

Touchscreen technology has gained increasing popularity over the last decade due to its natural and convenient human-technology interaction. It is widely used in a variety of personal, public and occupational settings [1, 2], including varied complex information systems implemented in healthcare facilities, aviation industry, and nuclear power stations [3, 4].

The use of touchscreen technology can bring a number of advantages. First, touchscreen can be easily accessed and operated by a wide range of users, including

inexperienced and disable user [5–8]. Second, touchscreen can reduce physical dimensions of a device, as physical keys that are usually applied in traditional input devices such as keyboards or mice can be replaced by on-screen virtual keys [7, 9, 10]. Third, it appears that the intuitive human-technology interaction provided by touchscreen can make the technology more attractive [6]. Finally, the design of touchscreen interface can be easily adjusted to provide only the keys that are relevant in a specific task at a given time (e.g., providing digit keys only in digit input tasks) [11].

In spite of its convenience and potential benefits, touchscreen interfaces, if poorly designed, are likely to result in frustration and irritation for users, in inefficiency and disruption in work process, and in a higher likelihood of committing errors [12, 13]. These negative outcomes might rise safety issues to both systems and users [7]. Therefore, it is important to investigate the effects of interface design factors that may affect the use of the technology so as to provide usable and safe touchscreen interfaces.

Previous studies have addressed several important design factors for touchscreen interfaces, such as key size and gap between keys [11, 14–19]. For example, Pfauth and Priest suggested that key size is one of the most important factors in touchscreen use when the touchscreen interface involved a hierarchical menu display [20]. Chen et al. investigated the effects of key size and gap size on touchscreen input performance by individuals with varied motor abilities [14]. Their results indicated that as key size increased, the performance for disable participants improved, while the performance for non-disable group plateaued at a key size of 20 mm. Chourasia et al. evaluated the effect of posture (i.e., sitting and standing) on touchscreen performance and touch characteristics during a digit entry touchscreen task among individuals with and without motor-control disabilities [15]. They found that standing affected touchscreen performance only at smaller key sizes (i.e., key sizes smaller than 20 mm) and led to greater exerted force and impulse compared with sitting. Colle and Hiszem investigated effects of key size (i.e., 10 mm, 15 mm, 20 mm and 25 mm) and gap size (i.e., 1 mm and 3 mm) on touchscreen numeric keypad performance [11]. They found that task input time was longer and error rate was higher for smaller key sizes, while the performance plateaued as key size increased up to 20 mm. Jin et al. examined a number of key sizes (i.e., 11.43 mm, 13.97 mm, 16.51 mm, 19.05 mm, 21.59 mm, and 24.13 mm) and gap size (i.e., 0 mm, 3.17 mm, 6.35 mm, 12.7 mm and 19.05 mm) for touchscreen interfaces with older adults [19]. Their results indicated that 19.05 mm size yielded the highest accuracy rate. In addition, these studies consistently found that gap size did not affect user performance [11, 14, 15, 19]. However, they did not consider scenarios where there is no gap between keys.

Although key size and gap between keys have been widely examined [14, 15], consensus on design guidelines for the two factors seems lacking [14]. For instance, International Organization for Standardization (ISO) recommends that the size of the touch-sensitive area should be at least equal to the breadth of the index finger distal joint for the ninety-fifth percentile male, while the Electronic Industries Association (EIA) recommends 19.05 mm as the minimal touch-sensitive size. Moreover, the America National Standards Institute (ANSI) suggests a minimal key size of 9.5 mm with a 3.2 mm gap, and states that key sizes larger than 22 mm lead to no performance improvement.

Besides key size and gap, there exist other important factors (e.g., the location of key characters on keys) that may affect the usability of touchscreen but have not been previously investigated. Theoretically, key characters could be presented in any location on the key area. In practice, the location of key characters vary in different types of keys. For example, it is widely encountered in traditional physical keyboards that key characters located in upper left area for letter keys, in upper left, central or lower right area for some function keys, and in upper or lower left for numeric keys. However, there seems to have few design guidelines to guide the design practice for the location of key characters. In addition, the location of key characters is likely to provide users for behavioral cues, especially in touchscreen input tasks. It is possible that users may hit the character on a key directly rather than other areas of the key when they are required to click the key. Hence, the location of key characters may affect a user's decision on the area that their fingertips would touch on the keys, and thereby affect the input accuracy. Moreover, users' hands and fingers could cover key characters when they touch the keys. The degree to which key characters are covered may be different as the characters are located in varied locations within the key area. For example, key characters in lower right could be fully covered when users' fingers get close to the touchscreen interface, while key characters in upper left could be well seem until users' fingers press the interface. Therefore, the location of key characters may have important impacts on the usability of touchscreen devices in input tasks; but this speculation requires further confirmation.

The purpose of this study was to examine the effects of key size, gap and the location of key characters on the usability of touchscreen devices. While many previous studies focused on small, mobile devices [21–23] and examined digit input tasks [14, 15], we based our study on a large touchscreen interface and examined letter input tasks.

2 Methods

2.1 Experimental Design

A three-factor ($4 \times 2 \times 5$), within-subjects design was employed in this study, with key size (i.e., 10 mm, 15 mm, 20 mm and 25 mm), gap (presence (i.e., 2 mm) and absence), and the location of key characters (upper left, upper right, central, lower left and lower right) serving as independent variables, and sets of usability metrics (i.e., task completion time, accuracy rate and user preference) serving as dependent variables. Task completion time referred to the amount of time a participant needed to complete a task. Accuracy rate was calculated as the proportion of correctly input characters in a task. User preference was assessed using a paper-based questionnaire that asked participants to choose their preferred button design from varied levels of the three examined factors.

2.2 Participants

Fourteen undergraduate students (8 male and 6 female), aged from 21 to 23 years, participated in this study. All participants were right-handed and reported having

normal or corrected to normal vision. The study protocol was approved by the institutional review board of Shenzhen University.

2.3 Materials and Tasks

A DELL All-In-One touchscreen computer (screen size: 23 in., resolution: 1600 × 900) used to present our experimental tasks. The touchscreen was tilted back with a 70-degree angle from the horizontal level, as suggested by previous studies [14, 15]. Task scenarios were created by Visual Studio 2012 in a Microsoft Foundation Classes operating system. The experimental interface contained a target box, an input box and an experimental keyboard (See Fig. 1 for an example) with square keys. White was applied to the background of the touchscreen keyboard interface, with grey for key characters. The task required the participants to enter a random six-letter string shown in the target box using the experimental keyboard.

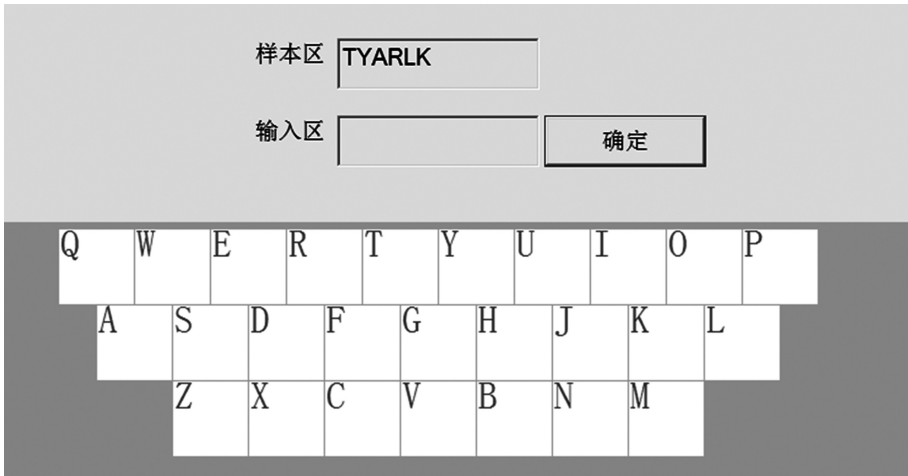


Fig. 1. Screenshot for an experimental touchscreen interface (Experimental condition: 20 mm key size, the absence of gap and key characters located in the upper left area. English words in parentheses are used for explanation only and would not show in the test).

2.4 Procedures

Before the experiment, participants provided informed consent and were given detailed information of test procedures. They were instructed to sit and adjust the chair according to their preference. Following several practice tasks to familiarize themselves with the test, participants were asked to click a start button on the center of the screen to initiate the main experimental tasks. Participants were asked to respond with their index fingers as quickly and accurately as possible. The combinations of key size, gap and the location of key characters were randomized in a full factorial design. There were three letter tasks for each of the combinations. Upon the completion of all tasks, the

preference questionnaire was administered to elicit participants' preference on button design. The whole experiment could be completed within half hour.

2.5 Data Analysis

Three-way repeated measures analyses of variance (ANOVAs) were used to determine the main and interaction effects of key size, gap and the location of key characters on user performance. Past hoc analyses were performed with Bonferroni adjustment where necessary. Chi-square test was performed to examine the difference in user preference. Level of significance was set at $\alpha = 0.05$. Statistical analyses were performed using SPSS Version 22.

3 Results

3.1 Task Completion Time

Table 1 presents ANOVA analysis results for task completion time. There was a significant main effect of key size on task completion time while gap and the location of key characters did not yield any effect. On average, the task completion decreased by 8% as the key size increased from 10 mm to 20 mm. The location of key characters had a marginal interaction effect with key size ($F(12, 156) = 1.729$, $p = 0.065$) (Fig. 2), but not with gap (Fig. 3).

Table 1. Effects of key size, gap and the location of key characters in task completion time (s).

	Descriptive analysis		ANOVA	
	Mean	SD	F values	p values
Key size			4.137	0.006
10 mm	5.9	1.5		
15 mm	5.5	1.2		
20 mm	5.4	1.6		
25 mm	5.7	1.4		
Gap			0.495	0.482
Absence	5.7	1.6		
Presence	5.6	1.3		
The location of key characters			0.675	0.61
Upper left	5.6	1.3		
Upper right	5.7	1.5		
Central	5.6	1.6		
Lower left	5.8	1.2		
Lower right	5.5	1.4		

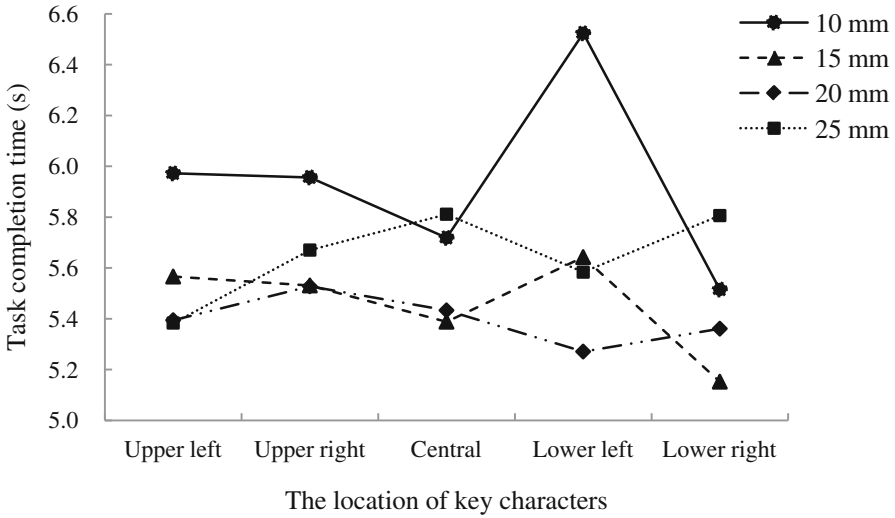


Fig. 2. Task completion time (s) by key size and the location of key characters.

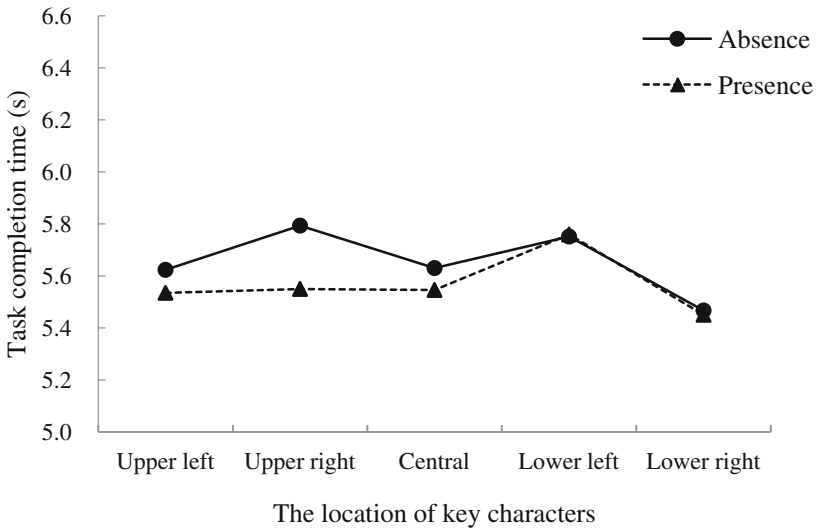


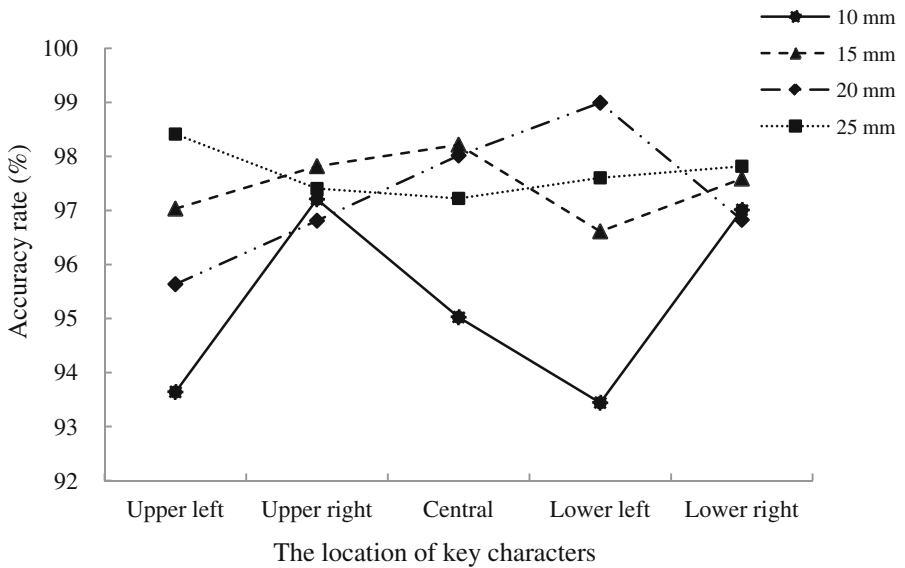
Fig. 3. Task completion time (s) by gap and the location of key characters.

3.2 Accuracy Rate

Table 2 presents ANOVA analysis results for accuracy rate. Key size was found to have a significant effect on accuracy rate while gap and the location of key characters alone did not yield any effect. On average, the accuracy rate increased from 95.2% to 97.6% as the key size increased from 10 mm to 25 mm. Accuracy rate plateaued at 15 mm key size with little improvement for larger key sizes. The location of key characters had a

Table 2. Effects of key size, gap and the location of key characters on accuracy rate (%).

	Descriptive analysis		ANOVA	
	Mean	SD	F values	p values
Key size			4.761	0.030
10 mm	95.2	8.3		
15 mm	97.4	4.9		
20 mm	97.2	6.0		
25 mm	97.6	4.5		
Gap			2.112	0.147
Absence	97.2	5.2		
Presence	96.5	7.0		
The location of key characters			0.723	0.576
Upper left	96.1	8.7		
Upper right	97.2	4.7		
Central	97.1	6.8		
Lower left	96.6	5.0		
Lower right	97.2	4.6		

**Fig. 4.** Accuracy rate (%) by key size and the location of key characters.

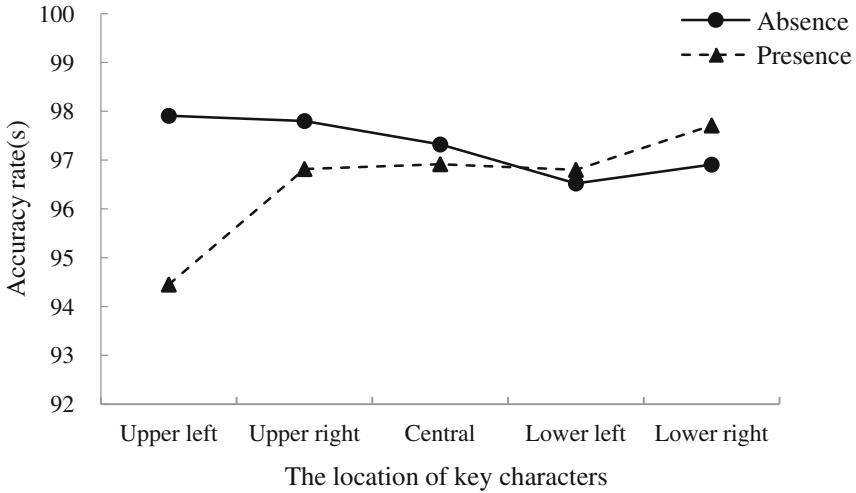


Fig. 5. Accuracy rate (%) by gap and the location of key characters.

significant interaction effect with key size ($F(4, 52) = 2.634, p = 0.044$) (Fig. 4), but not with gap (Fig. 5).

3.3 User Preference

Table 3 shows the user preference data on key size, gap and the location of key characters. The majority of participants preferred 15 mm key size (64%, $\chi^2 = 16.286$,

Table 3. Distribution of user preference by key size, gap and the location of key characters.

	Percentage
Key size	
10 mm	0%
15 mm	64%
20 mm	36%
25 mm	0%
Gap	
Absence	43%
Presence	57%
The location of key characters	
Upper left	29%
Upper right	0%
Central	71%
Lower left	0%
Lower right	0%

$p = 0.010$), the presence of gap (57%, $\chi^2 = 0.286$, $p = 0.900$), and centrally located key characters (71%, $\chi^2 = 27.429$, $p = 0.001$).

4 Discussion

This study examined the effects of three touchscreen design factors (i.e., key size, gap and the location of key characters) on user performance and preference during letter input tasks. In general, key size yielded a significant effect on touchscreen input performance, while gap and the location of key characters alone had no measurable effect. User performance improved as the key sizes increased up to 15 mm. The location of key characters was found to marginally interact with button size in terms of both task completion and accuracy rate. Users generally preferred 15 mm key size, the presence of gap between keys, and key characters that were centrally located.

4.1 Effects of Key Size

The results indicate that user performance improved as key size increased up to 15 mm. Little benefit could be obtained in larger key sizes. The value is within the recommended range of 15–20 mm for minimal usable touchscreen button sizes that are reported in previous studies [11, 14, 15, 18, 19]. However, it should be noted that the recommended range is relative broad, suggesting that variations exist, and the context of use should be considered in the touchscreen interface design. For example, Chourasia et al. found that while 15 mm button size was sufficient to achieve optimal performance during sitting, 20 mm button size was required during standing [15]. Chen et al. found that healthy adults could only get minimal gains from button sizes larger than 20 mm, while disabled adults continued to obtain benefits as button size increased above 20 mm [14].

4.2 Effects of Gap

Previous studies reported no effect of gap size on user performance [11, 14, 15, 19]. However, these studies only examined the presence of gap size (e.g., 3 and 5 mm in study by Chourasia et al. [15], and 1 and 3 mm in study by Chen et al. [14]), and ignored the absence condition (i.e., no gap between keys). Our study extended previous research by examining both presence (i.e., 2 mm) and absence of gap. Results indicated that the absence of gap size did not affect user performance. One explanation for this may be that the difference between the presence and absence of gap manipulated in our study was small so that no effect was detected. This finding has important implication in that gap between keys could be removed to obtain a minimal touchscreen keyboard area, especially in cases where the touchscreen interface is limited. However, it should be noted that users generally preferred touchscreen keyboards with the presence of gap.

4.3 Effects of the Location of Key Characters

Another important contribution of the present work to the literature is that we provided empirical evidence on the effects of the location of key characters. In particular, our study examined five commonly encountered locations of key characters (i.e., upper left, upper right, central, lower left and lower right). Although it is assumed that the location of key characters might provide behavioral cues for users and affect input performance, we found no effect of the location of key characters on touchscreen input performance. One reason for this could be that most of our touchscreen keys were large enough so that the participants are unlikely to hit the area outside the key scope. We found that the location of key characters interacted with key size in terms of accuracy rate. The accuracy rate was lower if the keys were located in the upper and lower left in the key area for the smallest key size (i.e., 10 mm). Another explanation may be that the key characters in our study were also large enough so that key characters could not be fully covered by users' hands and fingers wherever the key characters was located. Users could always see the characters in varied location conditions and obtain behavioral cues. However, it should be noted that participants generally preferred centrally located key characters, which could be applied in design practice to improve user satisfaction.

4.4 Implications and Future Directions

The results of this study lead to the following implications for the design of touchscreen interfaces. First, a key size of 15 mm could be recommended as the minimal usable option across gap and the location of key characters. However, this value came from healthy young adults in a static, sitting posture. Whether it could be an equally optimal size for other user groups and across varied contexts of use is unknown and requires further confirmation. Second, gap between keys, though yielded no measurable effect on user performance, could be an important design factor in touchscreen interface design. Removing the gap between keys could save space for limited touchscreen interfaces and would not lower user performance, while the presence of gap is likely to increase user satisfaction. Finally, our results did not indicate an optimal location of key characters in user performance. In this case, a centrally located key character could be a better option as it was preferred most by users. However, more research efforts are required to examine the effects of the location of key characters in relation to other factors, such as tasks and use scenarios. In general, specific design guidelines regarding the examined factors in our study are lacking and should be established from future empirical evidence.

5 Conclusions

This study examined the effects of key size, gap and the location of key characters on touchscreen use. In general, key size had a significant effect on touchscreen input performance. The performance improved as the key sizes increased and plateaued at 15 mm. The presence of gap and the location of key characters, though having no measurable effect alone, affected the use of touchscreen interface through their

interaction effects with key size and with each other. The factors examined in this study require further examination to determine their effects in a variety of touchscreen usage scenarios. The results may help with the design of more usable and safe touchscreen technology.

References

1. Gill, G.K., Shergill, G.S.: Perceptions of safety management and safety culture in the aviation industry in New Zealand. *J. Air Transp. Manag.* **10**(4), 231–237 (2004)
2. Tanaka, S.: Accident at the Fukushima Dai-ichi nuclear power stations of TEPCO. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **88**(9), 471–484 (2011)
3. Harris, J., Reinerman-Jones, L., Teo, G.: The impact of simulation display on nuclear power plant task error frequencies. In: Cetiner, S.M., Fechtelkotter, P., Legatt, M. (eds.) *Advances in Human Factors in Energy: Oil, Gas, Nuclear and Electric Power Industries*. AISC, vol. 495, pp. 133–144. Springer, Cham (2017). doi:[10.1007/978-3-319-41950-3_12](https://doi.org/10.1007/978-3-319-41950-3_12)
4. Or, C., Tao, D.: A 3-month randomized controlled pilot trial of a patient-centered, computer-based self-monitoring system for the care of type 2 diabetes mellitus and hypertension. *J. Med. Syst.* **40**(4), 81 (2016)
5. Ahearne, C.: Touch-screen technology usage in toddlers. *Arch. Dis. Child.* **101**(2), 181 (2016)
6. Neafsey, P.J., Strickler, Z., Shellman, J., Padula, A.T.: Delivering health information about self-medication to older adults: use of touchscreen-equipped notebook computers. *J. Gerontological Nurs.* **27**(11), 19–27 (2001)
7. Page, T.: Touchscreen mobile devices and older adults: a usability study. *Inderscience Enterprises Ltd.* **3**, 65–85 (2014)
8. Ng, H.C., Tao, D., Or, C.K.: Age differences in computer input device use: a comparison of touchscreen, trackball, and mouse. In: Rocha, Á., Correia, A., Wilson, T., Stroetmann, K. (eds.) *Advances in Information Systems and Technologies*. AISC, vol. 206, pp. 1015–1024. Springer, Cham (2017). doi:[10.1007/978-3-642-36981-0_96](https://doi.org/10.1007/978-3-642-36981-0_96)
9. Irwin, C.B., Sesto, M.E.: Performance and touch characteristics of disabled and non-disabled participants during a reciprocal tapping task using touch screen technology. *Appl. Ergon.* **43**(6), 1038–1043 (2012)
10. Parhi, P., Karlson, A.K., Bederson, B.B.: Target size study for one-handed thumb use on small touchscreen devices. In: *Conference on Human-Computer Interaction with Mobile Devices and Services* (2006)
11. Colle, H., Hiszem, K.: Standing at a kiosk: effects of key size and spacing on touch screen numeric keypad performance and user preference. *Ergonomics* **47**(13), 1406–1423 (2004)
12. Or, C., Tao, D.: Usability study of a computer-based self-management system for older adults with chronic diseases. *JMIR Res. Protoc.* **1**(2), e13 (2012)
13. Tao, D., Or, C.: A paper prototype usability study of a chronic disease self-management system for older adults. In: *IEEE International Conference on Industrial Engineering and Engineering Management* (2012)
14. Chen, K.B., Savage, A.B., Chourasia, A.O., Wiegmann, D.A., Sesto, M.E.: Touch screen performance by individuals with and without motor control disabilities. *Appl. Ergon.* **44**(2), 297–302 (2013)

15. Chourasia, A.O., Wiegmann, D.A., Chen, K.B., Irwin, C.B., Sesto, M.E.: Effect of sitting or standing on touch screen performance and touch characteristics. *Hum. Factors* **55**(4), 789–802 (2013)
16. Park, Y.S., Han, S.H., Park, J., Cho, Y.: Touch key design for target selection on a mobile phone. In: Conference on Human-computer Interaction with Mobile Devices and Services, pp. 423–426 (2008)
17. Pitts, M.J., Burnett, G., Skrypchuk, L., Wellings, T., Attridge, A., Williams, M.A.: Visual-haptic feedback interaction in automotive touchscreens. *Displays* **33**(1), 7–16 (2012)
18. Kim, H., Kwon, S., Heo, J., Lee, H., Chung, M.K.: The effect of touch-key size on the usability of in-vehicle information systems and driving safety during simulated driving. *Appl. Ergon.* **45**(3), 379–388 (2014)
19. Jin, Z.X., Plocher, T., Kiff, L.: Touch screen user interfaces for older adults: button size and spacing. In: Stephanidis, C. (ed.) UAHCI 2007. LNCS, vol. 4554, pp. 933–941. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-73279-2_104](https://doi.org/10.1007/978-3-540-73279-2_104)
20. Pfauth, M., Priest, J.: Person-computer interface using touch screen devices. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **25**, 500–504 (1981)
21. Conradi, J., Busch, O., Alexander, T.: Optimal touch button size for the use of mobile devices while walking. *Procedia Manufact.* **3**, 387–394 (2015)
22. Kwon, T., Na, S., Park, S.H.: Drag-and-type: a new method for typing with virtual keyboards on small touchscreens. *IEEE Trans. Consumer. Electron.* **60**(1), 458–459 (2014)
23. Shin, H., Lim, J.M., Lee, J.U., Lee, G., Kyung, K.U.: Effect of tactile feedback for button GUI on mobile touch devices. *ETRI J.* **36**(6), 979–987 (2014)

Natural, Multi-modal Interfaces for Unmanned Systems

Glenn Taylor^(✉)

Soar Technology, Ann Arbor, MI, USA
glenn@soartech.com

Abstract. The prospect of using unmanned systems in dull, dirty, or dangerous jobs to save work or even lives has drawn increasing attention in the DoD. Unmanned ground vehicles are used in theatre to get views into buildings or to destroy suspected IEDs. Unmanned air vehicles are used to get views over the next hill or to deliver munitions on targets thousands of miles away. While the automation and sensing capabilities have increased, interaction with these systems is still fairly rudimentary. Deployed systems typically use tele-operation or waypoint control, in some cases requiring operators to carry heavy operator control units. These approaches place a high burden on the operator in terms of the added weight and the constant attention required to operate the systems. In fact, many of these systems require more than a single operator to control a single platform, which increases the cost and logistics of using them. This paper describes natural, multi-modal interfaces as an alternative to the current state of the practice in controlling unmanned systems, with the goal of leveraging how people already communicate with each other in order to reduce the physical and cognitive burdens of interacting with unmanned systems. We describe approaches and challenges in designing, building, and evaluating natural interfaces. We present our Smart Interaction Device (SID) as an example natural interface for interaction with unmanned systems, and highlight some use cases we have applied it to in the air and ground domains.

Keywords: Unmanned systems · Natural interaction · Multi-modal interface · Dialogue systems

1 Introduction

The prospect of using unmanned systems in dull, dirty, or dangerous jobs to save work or even lives has drawn increasing attention in the DoD. Unmanned ground vehicles (UGVs) are used in theatre to get views into dangerous buildings or to destroy suspected improvised explosive devices. Unmanned air vehicles (UAVs) are used to get views over the next hill or to deliver munitions on targets thousands of miles away. While the platform and autonomy capabilities have increased, interaction with unmanned systems is still fairly rudimentary. (Throughout this paper, we will abbreviate ‘unmanned vehicle’ as UxV to generalize across different domains, and will use ‘unmanned vehicle’ and ‘unmanned system’ interchangeably.) Deployed ground systems use tele-operation for control, making them essentially remotely controlled vehicles with cameras and

actuators. Deployed UAVs typically include some autonomy for waypoint or route following, along with control loops to keep the aircraft aloft. Because the user's interaction with these systems is at a fairly low level (tele-operation or low-level commands to set waypoints), there is a high burden on the operator. In fact, many of these systems require more than a single operator to control a single platform, which increases the cost and logistics complexity of using the platforms. Much of this has to do with lack of autonomy – these deployed systems lack the capabilities to perceive or navigate their environment without running into problems very quickly. The result is that human operators are relied on to control much of UxV movement and sensing.

The DoD expects that advances in autonomy will allow unmanned ground vehicles to move along with a squad to carry their extra gear, or swarms of unmanned air vehicles to quickly search an area. Even further, the DoD is looking to have unmanned systems act as teammates rather than tools. Greater autonomy can help expand the uses of these systems, making the platforms more capable and reducing some burden on the operator, but there is still the need for an operator to communicate tasking or other information to the UxV in an efficient and effective manner. An operator who is performing other tasks will not be able to attend fully to one or more UxVs, either to drive or monitor them. To be more useful in operations, communicating with these systems needs to be as easy and as natural as communicating with a teammate or a subordinate.

This paper describes natural, multi-modal interfaces as an alternative to the current state of the practice in controlling unmanned systems, with the goal of leveraging how people already communicate with each other in order to reduce the physical and cognitive burdens of interacting with UxVs. We describe approaches and challenges in designing, building, and evaluating natural interfaces. We also describe our Smart Interaction Device (SID) as an example natural interface for interaction with unmanned systems, and highlight some use cases we have applied it to in the air and ground domains.

2 Natural Interaction

Instead of using devices such as joysticks, game controllers, or keyboards as ways to interact with robots, an alternative is to consider the ways in which people interact with each other. By allowing users to interact in ways that are familiar to them, without having to learn new devices, the expectation is that users should be able to learn how to interact with these systems more quickly, and that they would find the user interfaces more intuitive and familiar in general. In some cases, they may also free a user from having to carry additional equipment (such as an operator control unit or a laptop) to communicate with unmanned systems.

People use many modes naturally. Speech is obviously a common mode of interaction, and this includes the words people speak as well as the prosodic elements such as pitch and intonation as ways to carry meaning. Gestures, too, are quite natural to people, and can include movements of the hands, arms, fingers, and other body parts, and these motions or held shapes are meant to convey specific meaning. In military domains especially, drawing maps and sketching *on* maps are common and natural

ways to interact. Gaze, body language, and even the distance people stand from each other are ways in which people naturally communicate.

A person's task, and the environment in which that task is performed, often influences the modes of interaction. If the environment is noisy, or the task demands absolute quiet, then spoken communication may not be useful or desirable. If two participants cannot see each other, gesture is not effective. Resource limitations also play a role – if one participant has her hands full, then gesture will not be possible.

Having multiple modes to choose from can be useful in different ways. In some cases, the same information can be conveyed in different modes – e.g., saying goodbye or waving goodbye. This redundancy allows someone to choose how to communicate in the moment without loss of information. Redundancy also allows someone to use two modes simultaneously to make sure the message is received. On the other hand, not all modes are equally capable or as effective at conveying the same information. For example, giving a verbal description of an object may be less efficient than simply pointing at it. Often people will mix modes to communicate effectively – e.g., verbally saying a destination while also tracing a path on a map. This mixing of complementary modes can be more efficient than trying using just one [1].

Besides the mode of interaction (speech, gesture, etc.), another facet of natural human interaction is dialogue – communicating over time, with contributions from two or more participants. Dialogue can be used to ensure that the participants have a common ground for what's being communicated [2]. Dialogue can also be used to overcome failures in communication, where one person may ask for clarification if something is not clear, or if some information was missed because of a noisy environment. Dialogue also aids in efficient communication, where the dialogue itself serves as context for understanding references that someone might use in the conversation. A sign that dialogue is an aspect of natural interaction is that a participant can get frustrated when another participant fails to follow conversational rules [2, 3].

Naturalness in an interface is not defined solely in terms of the modes of interaction. A gesture-based interface that forces the user to place her arms in uncomfortable positions is not really natural. A speech-based interface for a robot that consists only of the verbal commands, *forward*, *turn-left*, *turn-right*, and *stop* might use a natural mode, but the language itself is possibly not very natural to the task. To make robot to do anything useful would take a great deal of effort and time. One goal of many human-system interfaces is *supervisory control*, which Sheridan describes as having three requirements: some level of *system autonomy*, user *situation awareness* of the system's behavior, and *high-level interaction* [4]. A robot that still requires low-level control will never let the user achieve supervisory control because she will be too busy driving it inch by inch.

3 Related Work

The last few years have seen a surge in natural interaction in commercial products. Speech-enabled assistants such as Amazon Echo®, Apple Siri®, Google Now®, and Microsoft Cortana® aim to make tasks such as playing music, finding directions, or searching the internet easier. While their speech recognition and language understanding

capabilities are impressive, they are still fairly limited in the kinds of interactions they support, and typically don't give more than one-shot answers to questions or requests. Spoken interfaces to robotic systems have appeared in research systems for a long time. Fairly simple speech interfaces for robotic systems have also appeared in consumer-oriented robots, including the recent Cozmo® from Anki.

Researchers have investigated gesture recognition for some time, with impressive work being done in domains such as carrier deck hand-and-arm signals [5] and American Sign Language [6]. The recent introduction of relatively inexpensive sensor systems such as Microsoft Kinect®, Leap Motion®, and Thalmic Labs' Myo® arm-band have spawned a new wave of interest in gesture recognition as an input modality. Gesture-based interfaces have also begun to make their way into other commercial products. Some of the augmented reality headsets that are now available, such as Microsoft HoloLens® and Atheer Air®, include a handful of simple hand gestures as ways to select objects or invoke menus, in part out of necessity since these systems do not come with typical devices like keyboard and mouse.

Body posture, facial expressions, and gaze as user interface controls largely remain in the realm of research labs. Likewise, while mixing modes such as speech and gesture together is a common trait of human interaction, it is still largely a research endeavor. There are some exceptions, of course. For example, gaze-based control has found a place in user interfaces for physically disabled users, using gaze to for a wide variety of computer tasks [7]. Microsoft's HoloLens® allows a user to move a cursor with his head (a proxy for gaze), and, when the cursor is over a button, the user can say "select" to press the button.

Dialogue is another dimension of natural human interaction, and a number of dialogue systems have been researched, from personal assistants that help make reservations [8] to chat-bots for question-answering [9]. While dialogue systems exist in commercial products, most are quite limited. They take the form of customer support systems, such as automated phone systems for airlines or banks that essentially a user through a menu via speech, or in the form online chat-based help systems that help screen questions before a human operator takes over.

Perhaps the closest to the work we describe here is that of the WITAS system [10], which includes mixed modes and simple dialogues for natural interaction with unmanned systems. Our work is also inspired by the QuickSet system [11, 12], in which users could engage in multi-modal dialogues to construct military simulation scenarios. Our work also squarely fits into the area of supervisory control [4], in which we aim to supplement the UxV's autonomy by raising the level of interaction with them, as well as helping the operator maintain awareness about the system's behavior.

4 Designing for Natural Interaction

As might be gleaned from the earlier discussion of natural interaction, how interaction happens in practice can be quite involved, and is affected by the task, the environment, the reliability of the communication channel, the skill of the participants, and their choices of how best to communicate. If we are to design natural user interfaces for people to interact with unmanned systems, we need to understand how natural

interaction happens in context, and provide for the kind of flexibility and resiliency that is afforded by natural interaction in those contexts.

4.1 Discovering Interaction in Context

Designing for natural interaction typically starts as a discovery process. In many cases, the UxV use case involves using the system in a similar role that a person would play (for example, a large UGV playing the role of rear security in a dismounted squad). In these cases, we can use the person-to-person analog as the model for how participants interact. If we limit the scope of interaction with unmanned platforms to task-oriented communication (as opposed to, for example, small talk), we can start by studying how people communicate naturally when performing the task. Standard task analysis methods can provide a framework for understanding the task itself and even the cognitive elements of the task [13], but these methods do not typically focus on communicative aspects of a task. Instead, methods such as interaction analysis [14] focus on the interaction itself. From the interaction perspective, we need to understand a number of elements. What language(s) do participants use to communicate? What modalities do they use, and under what conditions? What do they talk about? Why do they communicate? What do they do when communication breaks down? How does the context change the communication? Consider some examples. If a squad needed to maintain quiet while moving to occupy a position, how would the squad leader tell the squad to halt? If a team were geographically dispersed, how would a leader communicate a new mission to a remote subordinate unit? Understanding these elements is the first step in building a user interface when an unmanned system is one of the teammates.

In other cases, a UxV may be used in a new task, or offer a new capability, for which there is no real analogue in that use context, or for which there are only typical “artificial” interface equivalents. For example, dismounted infantry units might have access to small UAVs, but the only interface to them may be through joystick or GUI control on a laptop. If we want to introduce natural modes of interaction, we must discover how these users might *want* to interact with these systems.

Wizard of Oz (WoZ) studies are a standard method to discover new types of interactions and user preferences [15]. For example, suppose a user didn’t have access to a keyboard and mouse, but instead had a map, a pen and a microphone, how would that user *want* to interact an unmanned system for different tasks? In Wizard of Oz studies, user interface mockups are built but are not connected to a real system, and instead a “wizard behind the curtain” makes it seem like the user interface is having an effect. These studies let researchers design hypothetical interfaces to explore different ways of interacting without having to also implement the real system. One challenge in designing WoZ studies is getting users outside of their mentality of using a typical user interface, which might be a practiced way of interacting with the system. Additionally, standard graphical interfaces often provide visual hints for the kinds of inputs the system can accept, which helps the user along. Natural interfaces do not often provide these hints, and, in their absence, users can be at a loss for how to interact without some kind of prompting or training. Another challenge in Wizard of Oz studies is giving the

“wizard” all the right tools and interfaces to implement the expected behavior of the system; without good supporting tools, the wizard’s job can be quite frantic behind the curtain.

In these latter cases where there is no prescribed natural way of interacting with a UxV, it might also be useful to look for other analogs. For example, a typical UAV operator in a small infantry unit might be accustomed to flying the vehicle directly with a joystick in order to get the vehicle in position to see inspect an area. However, the unit leader likely gave a command to the UAV operator in some natural way. That operator’s job includes translating the unit leader’s commands into commands for the UAV using the joystick. To make the operator’s interaction more natural, we might look to the leader-operator interactions as a way to understand how the operator might want to interact with the vehicle. This may only be a starting point, however; the UAV operator brings particular skills in translating the leader’s commands into UAV commands, and there may be other types of inputs the operator would need to give to the system, or other types of feedback the operator would need from the system, for effective operation.

4.2 Designing for Failure

Accounting for the practical differences between truly natural interactions as humans perform them, and the limitations of today’s sensors and recognition systems, is one of the challenges of actually building these kinds of systems. For example, speech recognizers, while getting very good, are still not wholly reliable. Often one compromise is to limit the interaction language to account for these deficiencies – for example, by limiting the lexicon or grammar to a particular sub-language. However, this can have the negative effect of putting increased burden on the users to “unlearn” what might be natural for them, and then learn the limited inputs the system can actually recognize, with regular mistakes when they fall back to their typical language.

Another approach to this problem is to design for failure from the beginning. When two people communicate, one might fail to perceive an input, fail to recognize the input, or fail to understand the meaning of the input, for any number of reasons. Automatic recognition and understanding systems will also never be 100% reliable, and so the system must be designed in such a way as to accommodate these different types of failure.

Giving the user insight into what was recognized before the system acts on the input is one approach. Texting apps on smart phones often take this approach: when speaking instead of typing in a text window, the phone will often show the results of the speech recognizer before sending the text. This gives the user an opportunity to correct any errors manually before accepting the result. Of course, there is a cost – the message does not get sent out very quickly if the user reviews the message, and has to fix it, it beforehand.

Providing feedback to the user when recognition succeeds can also be helpful in maintaining the user’s awareness about what the system is doing. If the direct result of the input is not obvious (for example, the unmanned system is not visible), or the effect will not happen for some period of time, the user may have no way of knowing what

input was recognized or what command was issued. Making the user aware at least gives her the opportunity to make an adjustment. Whether the user corrects the recognition before the recognized input is acted on or after the user notices something went wrong in the system is akin to a *management by consent* versus *management by exception* approach to supervisory control [16].

This simple input-with-feedback exchange is an example of a type of dialogue with the user, with the intent of giving the user some awareness of what the system is doing. In addition to helping keep participants in sync, dialogue can help overcome failures in communication. Unlike the texting example above, there is no chance in regular spoken interaction for a person to see what another participant will recognize or understand. Dialogues might include asking the speaker to repeat when something was not heard, or asking for clarification when something is not understood. These different types of dialogue give the participants a chance to recover when something goes wrong. In building natural interfaces for unmanned systems, giving the system the ability to exercise these dialogue strategies can help keep the user aware of the system is doing as well as help overcome the inevitable failures in communication.

5 Smart Interaction Device

Over the last several years, we have been developing a system we call the Smart Interaction Device (SID), a natural multi-modal, dialogue-based interface to help users interact with unmanned systems [17, 18]. SID acts as a facilitator between a user and unmanned systems, allowing the user to communicate in familiar terms, and providing task-relevant information back to the user. SID allows for multiple input modes, including redundant modes and complementary modes, and communicates in multiple modes back to the user. SID frames all interaction with the user as a dialogue, working within familiar dialogue protocols to facilitate shared understanding and to overcome communication failures, and using the dialogue itself as context for understanding the user's inputs.

Figure 1 shows a high-level depiction of SID's architecture. SID Core contains three domain-independent modules that help facilitate the user's interaction with the unmanned platforms. The Dialogue Management component maintains the dialogue state and uses it as context to understand user inputs and provide feedback to the user. The Situation Awareness component maintains awareness of the unmanned platforms (progress and problems) and generates alerts and notifications in user terms rather than platform terms. The Planning and Execution module translates user-level commands into commands that the unmanned systems can understand. The processes within these modules can be tailored to specific tasks and domains.

Two implementation-specific modules are also part of the architecture. On the left side of SID Core is the Multi-Modal I/O component that fuses together multi-modal user inputs and also generates multi-modal outputs. Between the user and the system (and not depicted) is a range of input devices, sensors, and recognizers that take the raw inputs and generate semantic parses from them. Depending on the use case, these include recognizers for speech, gesture, and sketch, along with associated language parsers, the outputs of which are passed to the input fusion process, which generates

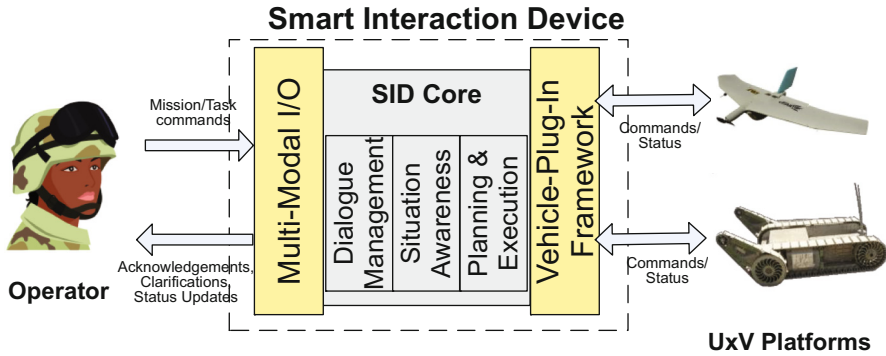


Fig. 1. High-level architecture of the Smart Interaction Device

hypotheses about the user’s intent. Likewise, there are a variety of output devices and displays that portray information to the user: speech generation, video, text, graphics, and haptic devices. On the right side of SID Core is a plugin framework for interfacing directly with a range of unmanned systems.

An important aspect of SID, and other user interfaces we have built for unmanned systems, is providing transparency into the behavior of the UxV. This relies in part on the systems themselves providing information back about their progress and status, which SID then translates into user terms. In some cases, there may be standard reports that a user expects and which is built into the domain-specific application – e.g., the robot reached a destination or spotted a threat. SID also allows the user to query status on demand, such as by asking “Robot, what are you doing?” or “Where are you going?” This information may also be on a graphical display, but it may also be easier for the user to ask for status verbally rather than look at a screen. Providing natural ways of querying the system and conveying information back to the user in different ways is also important to keep the user informed throughout the varying situations.

The architecture depicted above describes a general framework that we have instantiated for several use cases. The core behavior of the system has largely remained intact, but we have extended it to apply to a particular domain and task. Likewise, the specific I/O devices and platforms typically vary per domain and task, so we have built particular adapters to for these devices and platforms. In this section, we describe two different applications of SID, which include different use cases, different I/O devices, and different interaction languages.

5.1 Remote Operations

There are many use cases in which the operator and the unmanned systems may not be co-located, and in fact may be tens or thousands of miles away from each other. Even many small UAV operations are beyond line of sight, where the operator needs to rely on some external means to get a sense of the vehicle location and current task. In these types of operations, a map is a standard tool that people use to coordinate behavior.

Where both participants have an identical map that each can see and refer to, they use the map for coordination and situation awareness. The map can provide visual landmarks that can be used as reference points for tasking (“Fly to the bend in the river”) or for reporting location (“I just reached the bend in the river”). Map-based tasks are especially rich with mixed-mode interactions [19] where, for example, one participant can speak some information while drawing on the map to convey other information (e.g., saying, “Go here” while pointing to the map).

We have implemented in SID these kinds of interactions using a tablet-based map to show UxV positions and tasks, and to allow the user to create new reference objects (e.g., waypoints, routes) and tasks that refer to them (e.g., “Follow this route” while sketching the route). Map-based interactions such as drawing or pointing are through touch on the screen. Speech is through a microphone on the tablet or worn by the user. Once these reference objects are created and named, the user can refer to them later – e.g., “Follow route blue.” A Wizard of Oz study early in the design process helped us identify the kinds of interactions users would want supported with this kind of system [18]. The study showed that pointing gestures as well simple sketches of points, routes and areas could account for over around 85% of the drawn inputs. This allowed us to build an effective interface using only a few simple drawing elements.

An example of one such implementation is shown in Fig. 2 below. On the upper left side is a log of the conversation between the user and the system, which gives the user a sense of what they have been conversing about. In this case, we also display what was recognized last, and the user must deliberately send the message after checking it, with the ability to manually change it if something was misrecognized. On the right side is a display that shows the position of the aircraft on the map (labeled as STX-1), along with a waypoint and two areas that have been constructed by the user.

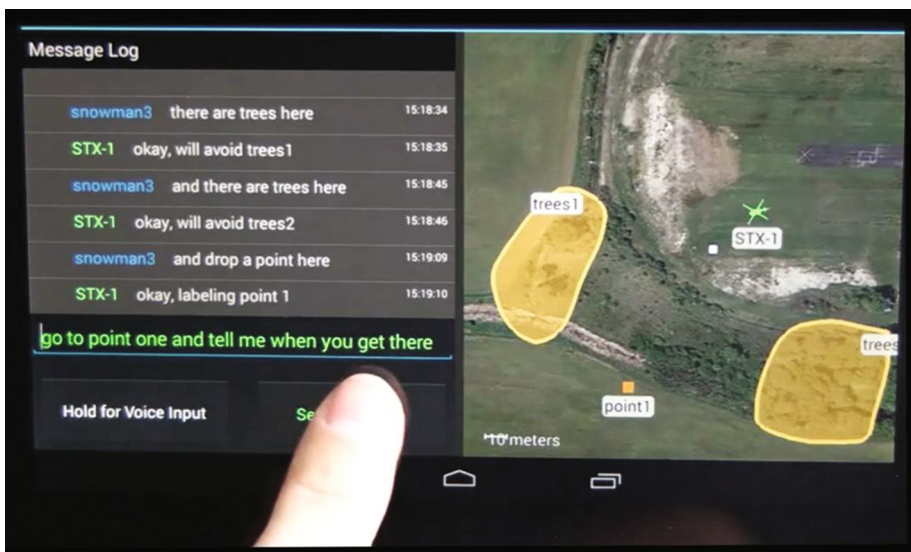


Fig. 2. A user interaction with a UAV with a speech-and-sketch map-based interface

Dialogue plays a role when the user gives ambiguous commands, refers to reference objects that don't exist, or when recognition fails for some reason. If the system failed to recognize the input, it will generically ask the user to repeat. If the user was ambiguous in the input, the system will ask for the specific information that helps complete the command (e.g., "Where do you want me to go?"). The user can reply with a brief answer to fill in the missing details, even using multiple modes in the reply (e.g., "Here" while pointing on the map).

5.2 Proximal Operations

Another use case is when the operator and the UxV are in close proximity such that they can see each other. An analogous use case in human operations would be a dismounted infantry squad on patrol. In this case, a large "robotic mule" UGV might travel with the squad to help carry extra equipment or to provide some surveillance capabilities. While a map-based interface such as described earlier could be used, one goal is to minimize the equipment that the operator has to carry and to keep his hands free for other tasks. Another goal is to help the warfighter keep his head up, looking at his surroundings instead of at a screen. Speech and gesture as modes help afford this kind of hands-free, head-up interaction, both among the members of the unit and between the operator and the UGV. In some cases, both speech and gesture are redundant modes: a command might be given using speech or gesture, such as in a "stop" command. The kinds of hand and arm signals in this domain are meant to be seen from a distance, so are fairly gross in detail and thus cannot convey a great deal of information (this is in contrast to, for example, American Sign Language, which is a complete language in and of itself, but is generally meant for face-to-face interaction).

We have implemented in SID these interactions using a few different technologies, most recently using a worn device as a means to capture both spoken inputs using a microphone and gesture inputs using the on-board inertial measurement unit (IMU) (see [20] for details). Redundant speech and gesture inputs cover basic tele-operation kinds of behaviors (forward, backward, left, right, stop), but also more autonomous behaviors such as following the user. Other speech commands that do not have gesture equivalents include taking control of the vehicle, releasing control, defining named rally points, and routes based on the position or movement of the UGV, maneuvering to those points or along those routes, and maneuvering in finer granularity such as turning a specific amount. Figure 3 shows some gesture interactions with a large UGV.



Fig. 3. Examples of gestures being given to control a large UGV (person on board the vehicle is a safety operator.)

6 Evaluation Approaches

There are many different approaches to user interface evaluations, most of which apply just as well to human-robot interfaces [21]. The inclusion of natural interaction as part of a user interface introduces new facets to the evaluation, and a few new challenges.

One factor that makes natural interfaces unique is that they can fail to recognize or understand the user. It is therefore important to measure the performance of input recognizers and related components in a wide range of conditions. Some of these conditions are environmental. For example, speech recognition is susceptible to surrounding noise, and even different types of noise may affect recognition differently (e.g., constant white noise, versus periodic, versus other people talking in the background). Different lighting conditions or other objects in the scene can affect vision-based gesture recognition. It is also important to evaluate the system with a wide range of users – ideally those who are good proxies for the target user. Speech recognizers may perform differently with different voices (e.g., from gender or age of the user, or accents), or with simultaneous user activity (e.g., running while speaking), or emotional state. Gesture recognizers have to account for different body shapes and sizes and different ways in which people might make a gesture. Each individual recognizer can be evaluated in isolation apart from other system-wide evaluation. While it does not tell the whole story about the user interface, it is an important first step to get a sense of how the system will perform and how natural the interface is.

User interface evaluations in general are often comparative, assessing one against another, and natural interfaces are no different. In comparative evaluations, learnability is often an important metric: how quickly someone can become a proficient user. One motivation behind natural interfaces is that they are meant to be closer to how people communicate, so should be easier to learn. Such hypotheses need to be tested for particular interfaces. If the specific inputs are artificial constructs that use natural modes but where the input language is invented or limited in some way that requires training, then learnability may suffer. For example, speech interfaces often have a fixed input language that the user has to get just right, otherwise the system fails to recognize the inputs. (Frequent and common deviations from this accepted language can be an opportunity to learn how the language coverage needs to be expanded.)

Other typical comparative metrics include the amount of time it takes to perform a task, and the amount of work for each task, in each of the two systems. In terms of natural interfaces, where recognition can fail, it is important to capture the cost of that recognition failure, both in terms of total time and in the amount of work (and frustration) it causes the user. Standard measures like NASA TLX [22] can be used to measure perceived workload, and progress on secondary tasks can serve as a proxy for objective workload. As with human communication failures, this failure can happen at multiple levels – in recognizing the raw inputs or in understanding the user in context – and understanding how and why this failure happens is important. It may also be important to measure how aware the user is of system failures; some failures may not be recognized, which can cause more problems further down the line.

Dialogue as part of a natural interface is also an important element to consider in the evaluation, and multiple approaches have been used to evaluate the utility, impact,

or even the naturalness of dialogue [23]. Where dialogue is used to help keep the user informed, it is important to measure that, such as by using standard situation awareness measures [24, 25]. Where dialogue is used to help recover from errors, it is important to measure the effect of the system on that recovery. Dialogue can also help make the interaction more efficient, leveraging the dialogue as context to help understand brief or otherwise ambiguous inputs. It can also be important to evaluate different dialogue strategies that might engage the user in different ways. In these cases, the study conditions would include isolating different strategies to measure their effect independently.

We will be exercising many of these aspects of evaluation in an upcoming study evaluating an interface for large robotic mules, with the help of the Army Research Laboratory's Human Research and Engineering Directorate (ARL/HRED) Field Element, and the participation of Soldiers at Fort Benning. There are several goals to the study. One is to look at user preferences between gesture and speech in cases where the input modes are redundant. We plan to put the interface through three trials on pre-defined courses: in one case, participants will use speech only; in another, gesture only; and in the third, they can use either mode interchangeably. We aim to understand not overall preference for a mode, but the contexts in which one is preferred over the other. Another goal is to assess the utility of some inputs that only have speech forms, such as defining routes and rally points, and tasking the vehicle using those control measures. In addition to quantitative measures of the effectiveness of different interactions, we also simply want to understand how Soldiers use the system in practice, based on their training and operational experience. This includes looking at the coverage of interactions we have so far and how they support Soldier tasks. This also includes understanding how the particular fit and form of the interface (a worn device) works with the way in which they operate. Putting the system in the hands of representative users and getting their feedback from an operational perspective is a critical step in understanding the interface's true value to its intended users.

7 Conclusions

Continuously improving recognition algorithms and the low cost of sensors has driven an increased interest in natural interfaces for computing systems, including for unmanned systems. In this paper, we have described an approach to designing and building natural interfaces for unmanned systems, including how to identify natural interaction in specific use contexts: how people interact today in these cases, or how they might want to interact with unmanned systems in the absence of traditional input devices. We have identified different dimensions of natural interaction, including the use of different modes (e.g., speech and gesture), ways in which use of these modes might be redundant or complementary, and ways in which interactions can become dialogues.

Natural interfaces can help improve the usability of unmanned platforms; however, this requires not just adding speech or sketch to the interface, but also taking into account the context of use. These modes have to be applied in ways that people find natural to a particular task, especially in ways in which they use those modes for the

same or similar tasks. In designing these systems, it is important to keep in mind how limitations in recognition technology can impact the design of the interaction languages themselves, and how these designs can subsequently affect the usability of the system. If the language understood by the system is too contrived or outside of what people use in their daily activity, the “naturalness” of the interface may be compromised and, in fact, make the system harder to use. It is also important to recognize that natural interfaces by – by the nature of natural inputs – will never be completely reliable, and to design the system to account for and recover from those failures. Framing interaction as a dialogue between the user and the system, and implementing dialogue strategies that are familiar to people, is one way to overcome these failures.

Natural interfaces also present some unique challenges for evaluation, for a number of reasons. Since one goal of natural interfaces is to be closer to how people communicate, it is important to design studies to measure this specifically. Because natural interfaces can fail to recognize or understand user input, it is also important to measure those failure rates and the impact they have on the task. Where additional mechanisms are used to mitigate those potential failures, such as shrinking the interaction language artificially or through the use of dialogue, the impact of those mechanisms must also be evaluated.

To illustrate some of these aspects of natural interfaces, from design to evaluation, we have described an example system, the Smart Interaction Device (SID) that incorporates many of the natural interaction features described. In the development of SID, we have applied it to multiple different domains and tasks, which has included integration with many different input devices and recognizers, covering speech, sketch, and gesture. We have also integrated and demonstrated SID with over a dozen different unmanned platforms, both air and ground, including heterogeneous, multi-platform demonstrations.

References

1. Oviatt, S., Coulston, R., Lunsford, R.: When Do We interact multimodally? Cognitive load and multimodal communication patterns. In: ICMI 2004. ACM, State College (2004)
2. Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)
3. Grice, P.: Logic and conversation. In: Morgan, J. (ed.) Syntax and Semantics. Academic Press, New York (1975)
4. Sheridan, T.B.: Telerobotics, Automation and Human Supervisory Control. MIT Press, Cambridge (1992)
5. Song, Y., Demirdjian, D., Davis, R.: Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011) (2011)
6. Pansare, J., Gawande, S., Ingle, M.: Real-time static hand gesture recognition for American sign language in complex background. *J. Signal Inf. Process.* **3**, 364–367 (2012)
7. Hutchinson, T.E., et al.: Human computer interaction using eye-gaze input. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **19**, 1527–1534 (1989)

8. Traum, D.R.: Conversational agency: the TRAINS-93 dialogue manager. In: Twente Workshop on Language Technology 11: Dialogue Management in Natural Language Systems (1996)
9. Quarteroni, S., Manandhar, S.: A chat-bot based interactive question answering system. In: Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue, Rovereto, Italy (2007)
10. Lemon, O., et al.: A multi-modal dialogue system for human-robot conversation. In: NAACL (2001)
11. Cohen, P.R., et al.: Quickset: multimodal interaction for distributed applications. In: 5th ACM International Conference on Multi Media. ACM Press (1997)
12. Johnson, M., et al.: Unification-based multimodal integration. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. ACL (1998)
13. Kirwan, B., Ainsworth, L.K. (eds.) A Guide to Task Analysis. Taylor & Francis, London (1992)
14. Jordan, B., Henderson, A.: Interaction analysis: foundations and practice. *J. Learn. Sci.* **4**(1), 39–103 (1995)
15. Kelley, J.F.: An iterative design methodology for user-friendly natural-language office information applications. *ACM Trans. Off. Inf. Syst.* **2**, 26–41 (1984)
16. Ruff, H.A., Narayanan, S., Draper, M.H.: Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence* **11**, 335–351 (2002)
17. Taylor, G., et al.: A multi-modal intelligent user interface for supervisory control of unmanned platforms. In: Collaboration Technologies and Systems Collaborative Robots and Human Robot Interaction Workshop, Denver, CO (2012)
18. Taylor, G., et al.: Multi-modal interaction for UAS control. In: SPIE.DSS, Baltimore, MD, April 2015
19. Cohen, P., McGee, D., Clow, J.: The efficiency of multimodal interaction for a map-based task. In: Applied Natural Language Processing Conference (2000)
20. Taylor, G., et al.: Multi-modal interaction for robotic mules. In: SPIE Defense and Security: Unmanned Systems Technology XIX, Anaheim, CA (2017)
21. Yanco, H.A., Drury, J., Scholtz, J.: Beyond usability evaluation: analysis of human robot interaction at a major robotics competition. *J. Hum. Comput. Interact.* **19**(1), 117–149 (2004)
22. 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.) *Human Mental Workload*. North Holland, Amsterdam (1988)
23. Paek, T.: Empirical methods for evaluating dialog systems. In: Workshop on Evaluation for Language and Dialogue Systems. ACM (2001)
24. Taylor, R.M.: Situational awareness rating technique (SART): the development of a tool for aircrew systems design. In: *Situational Awareness in Aerospace Operations (AGARD-CP-478)*. NATO-AGARD, Neuilly Sur Seine (1990)
25. Endsley, M.R., Garland, D.J. (eds.) *Situation Awareness Analysis and Measurement*, p. 383. Lawrence Erlbaum Associates, Mahwah (2000)

UI-Design and Evaluation for Human-Robot-Teaming in Infantry Platoons

Martin Westhoven^(✉), Christian Lassen, Irmtrud Trautwein,
Thomas Remmersmann, and Bernd Brüggemann

Fraunhofer Institute for Communication,
Information Processing and Ergonomics FKIE,
Zanderstraße 5, 53177 Bonn, Germany
martin.westhoven@fkie.fraunhofer.de

Abstract. The military benefit of unmanned reconnaissance for infantry as the most exposed military branch is obvious. Furthermore, unmanned systems can also support by transporting heavy equipment, including sensor payloads usually not fielded by infantry units. While larger assets are typically controlled from afar, smaller assets can be controlled directly by nearby troops and satisfy immediate reconnaissance needs. In this work, the design, implementation and evaluation of a user interface (UI) for integrating unmanned platforms into the German army's infantry platoons is presented. More specific, two unmanned aerial vehicles and two unmanned ground vehicles were to be integrated into a platoon. This work highlights the user interface aspects, training effort and organizational changes. German paratroopers and mountain infantry assisted with the requirements analysis and UI evaluation. In addition, the German Army Concepts and Capabilities Development Center supported the evaluation. The effort to bring unmanned systems into infantry units is motivated, related work concerning the control of unmanned systems is presented, the results of the requirements elicitation for this undertaking is reported, the design and implementation as well as the instruction strategy are outlined and the results of a test campaign reported. The paper concludes by summing up the current state and outlining future work regarding UI development for soldier-multi-robot-teams.

Keywords: Automation and autonomous systems · Command and control · Display design · Team working

1 Introduction

During dismounted operations, soldiers are most exposed to the enemy. Good reconnaissance can greatly improve survivability by enabling soldiers to avoid disadvantageous situations. Unmanned systems (US) can satisfy reconnaissance needs without risking human scouts. Furthermore, unmanned ground systems (UGV) can help lighten the typically high load for the individual infantrymen and therefore increase patrol range or provide transport for heavier equipment than infantry carries normally [1]. Small-scale US and those operating in close vicinity of dismounted troops can and

should be controlled directly as opposed to operation by a remote ground station. This eliminates additional communication outside of the platoon structure. To this end, the integration of two unmanned aerial vehicles (UAV) and two UGV into German infantry platoons was investigated, which included the whole process chain [2]. Aside from the user interface (UI) aspects, this included the development of the high-level communication protocol among others [3]. Controlling US can be cumbersome [4], even more so for dismounted units [5, 6]. If equipment is too complex, the operators need to set-up and their position has to be secured. This leads to reduced overall mobility of the platoon. Also, training time rises with increasing complexity and new structures or processes in the platoon. On the other side, the dismounted operation of US requires attention and thus reduces situational awareness [6]. Obviously, integrating robots into an infantry platoon is therefore a delicate process with many trade-offs to consider. In this work, we highlight the UI aspects of the overall project. We will first provide a brief overview over related work. The method chapter will go through requirements, the relevant hardware, the UI-design and the role concept. A description of the instruction for the soldiers follows and results from an evaluation campaign are presented and discussed. We conclude by summing up and giving an outlook to upcoming research and open questions.

2 Related Work

The presented work is based on previous research in the areas of autonomous air- and ground robot teams [7], human-multi-robot interaction with quasi-natural command languages [8] and the control of heterogeneous teams of robots [9]. Mi and Yang compiled a survey on human-robot interaction in swarms [10]. An important conclusion drawn from different sources is that the level of automation is case dependent. Designing for a human-oriented semi-autonomy fulfills best the requirements of keeping the user in the decision loop while automating non-critical tasks. Furthermore, trust in intermediate automation is higher than trust in fully automated systems [11]. Naturally, higher automation means less interaction required [12]. The balance between automation and human decision making therefore has to be designed carefully. Barnes et al. [13] give a broad overview on work by Israeli and U.S. researchers concerning the HCI design in context of dismounted control tasks for US. Reviews on Human Factors issues in US control form the base of this research [14, 15]. Some of the work also focuses the design for mixed initiative between human and autonomous decisions and is expanded by Barnes et al. [16]. A reduction of the required number of operators for a small UAV is reported by Stroumtsos et al. [17]. The resulting UI requires the user's full attention, since the waypoint navigation is relatively fine-grained. There is also work on the influence of agent transparency on operator performance that is the degree to which an operator is able to assess the current status and intentions of the system [18]. This is relevant for the UI design, as creating such transparency usually involves parts of the UI. Clare et al. [19] report on three different so-called Human Supervisory Systems (HSC) to control US. In another paper, Clare et al. report on scheduling strategies for mixed initiative control of a multi UAV system [20]. How to keep human operators in the loop, especially for making critical decisions, is investigated by

Franchi et al. [21]. They present a framework, which allows shared control between human and US as well as between multiple operators. Cook et al. [22] investigate attention guidance in such supervisory control settings. In Taylor et al., a multimodal supervisory control implementation is presented, which allows for different kinds of in- and output. Most importantly it aims to reduce misunderstandings between man and machine by dialogue-based interaction including questions to clarify the user's intent [23]. Endsley presents the human-autonomy systems oversight model, which aims among others at mitigating the decrease of operator awareness when autonomy is rising [24]. Dawson et al. [25] highlight cooperative supervision, which plays a part in this work as well. Demir et al. [26] focus on a similar topic, namely the team's situation awareness while cooperatively supervising US.

Evans [27] reports benefits of displaying planning information of US to the supervisor, which show themselves in reduced direct tele-operation. A mobile ground control unit on a tablet for search and rescue is used in Peschel et al. [28], but it is used to control only one UAV. They conclude that dedicated mission specialist interfaces can benefit role performance if designed and implemented carefully. Pitman and Cummings describe the design of hand-held touch controls for a single UAV with a focus on manual controls [29]. Their work has been extended to include object detection warnings and collision alerts [30]. An ongoing research project is the Unmanned Tactical Autonomous Control and Collaboration (UTACC) effort of the U. S. Marine Corps Warfighting Laboratory (MCWL) [31]. It focuses on cognitive load and how to enable warfighters to better cope with the ever rising information flood achieved by technical systems on modern battlefields. Bommer [32] also studies mental workload for multi-vehicle control and concludes that reducing task complexity is the key to enabling operators to control large amounts of UAV.

3 Method

In the following we present and motivate underlying requirements and how they were compiled, relevant hardware used in the system and the UI-Implementation itself.

3.1 Requirements

The problems of integrating US into infantry units result in conflicting requirements. Aerial reconnaissance experts and instructors of the German paratroopers were therefore interviewed to set priorities. First of all, required attention was reduced by widely automating standard operations and providing high-level controls. More details are required to access all system functionality, while realizing quick access. The mission context requires taking interruptibility, consistency and also environmental lighting into account. Organization-wise, integration into the existing structure of platoons is mandatory. Requirements also follow from the intended system function and some were set beforehand by the procurement bureau. Basic requirements include controls for the swarm and sub-groups. Natural sub-groups are ground and air systems and go down to single systems. Commands are annotatable with the intentions defensive,

neutral and aggressive where applicable. The platoon leader is able to adjust the distance of vanguard robots. UGVs are assignable to squad leaders and to follow them. UGVs can integrate into defensive positions while UAVs support reconnaissance and stationary sensors can be added. UGVs evade together with the soldiers and keep out of firing lines to avoid ricochets. UAVs report low battery status and await approval for a battery swap, except in emergencies. Feedback is automatically prioritized and acquired from the user. UGVs perform automated transport tasks between two points with pauses to handle freight. UGVs can follow a person. Platoon formation can be dissolved, resulting in execution of individual commands. US report the following objects and events: nuclear, chemical, biological and radiological agents (NCBR), improvised explosive devices (IED) and mines, jamming, snipers, alarm posts. Further required functions are: Move to coordinate, guard coordinate, observe coordinate, provide imagery, reconnoiter an area, cancel order, return to base or rallying point, emergency stop and manual control for UGVs. Additionally displayed items are the system and equipment status, completed commands, acquired imagery. System status includes the current task, if in formation, overall task, damages, battery, connectivity and stock of deployable equipment. These requirements are also partially found in other US efforts, e.g. in the U.S. UGV Interoperability Profile [33].

Infantry operations require high attention, resulting in a need for quick interaction, but also the consideration of interruptions. Consistent user guidance helps intuitive use. This includes menu design, map overview and all the icons and interaction-elements used. Especially displays in smartphone-size require optimal use of the available display space. Outdoor use necessitates good readability regarding contrast and luminance. To include the user's point of view, two semi-structured interviews were performed with two aerial reconnaissance experts and two paratrooper platoon and squad leaders. Furthermore, eleven instructors from the paratroopers were polled with a questionnaire. The interview included closed and open questions. Open questions served to poll facts and circumstances as well as a view into user motives. After presenting the study, the current procedures for a specified scenario were queried: An infantry unit redeploys to a new emplacement. Terrain, roads and buildings are known, but obstructions by enemy forces are to be expected, e.g. ambushes, IEDs or blockages. The focus was on infantry leadership and reconnaissance. Questions targeted what an appropriate course of action in this situation is, how reconnaissance is reported and what consequences follow from these reports. After the functions provided by US were explained, it was queried how these systems would be used and which information and input would be required by the soldiers. Questions targeted the tasks US would perform, the required user input, environmental conditions, operator tasks and crucial feedback. A qualitative analysis was performed and a questionnaire for validation and prioritization developed. Priorities were high, medium or low.

All topics could be expanded by own ideas. Specific topics were the event types to report, information types to display on the map, helpful map overlays, preferred US controls and general issues, e.g. storage in troop carriers. Reports of enemy units and IEDs are of high priority. A possibility for manual classification by type, strength and behavior of units and to mark recognized objects in video feeds was highly prioritized. Terrain information and information on neutral units and crowds were a medium priority. Additional mentions were unit movement directions, own forces, motivation

and behavior of forces. High priority information to be displayed on the map were own, allied and enemy positions, IEDs, terrain information, satellite image, tactical icons, north-arrow and position or direction of enemy shooters. Neutral units and sensor orientation were ranked medium and air-space order as low priority. Additional mentions were a 3D-view and a Universal Transverse Mercator (UTM) grid. Of high priority for US control was waypoint-navigation, observation areas and backup controls. Of medium priority was the definition of combat sectors. There were highly prioritized general concerns about US reliability and handling, namely US storage in troop carriers, terrain traversability, transport and endurance of batteries, specifically for UAV. In the discussion about map overlays, only single ideas were posed in the questionnaire regarding what would be useful. Mentions were air-space order, color-filters, blockages, terrain type, operation plan, event history, road network, night-vision, reconnaissance imagery, waypoints, own units, ethnics and enemy.

3.2 Organizational Integration

An important part of the integration into the existing structure of an infantry platoon is fitting the US into the existing command hierarchy. While the platoon leader has to decide in which way the US should be used, his or her other tasks already generate a significant load. Permanent direct control by the platoon leader therefore cannot be realized. This problem was discussed during the interviews and a new assistant role is proposed. The squad leader unmanned systems, or UXS, is to function analogous to the other squad leaders, as far as the platoon leader is concerned, but commands robots instead of men himself. Like the platoon's radio operator, he is integrated into the so-called "Zugtrupp", the team directly surrounding the platoon leader and normally consisting of the platoon leader himself, a radio operator and two additional soldiers for protection. Thus, the platoon leader can have direct access to the UI and the squad leader UXS is protected in situations where more complex interaction is necessary. The latter can possibly mitigate the generally greater vulnerability of dismounted units, which seems to affect the subjective workload while controlling US [4]. This allows for the use of a tablet-sized device, which would be prohibitively large for platoon members operating without dedicated protection.

The squad leader UXS is furthermore assisted by an artificial intelligence, the so-called swarm control. It mainly performs the translation of high level to low-level commands and manages the US in holding formation with the human part of the platoon. This solution also roughly follows results reported by Oron-Gilad and Parmet [34], which indicate a reduced attention to surroundings and higher required training for continuous or repeated dismounted use of video feeds from UGV. To ease information sharing in the platoon, the remaining squad leaders are fitted with smartphone-sized devices in forearm-mounts to mainly consume information, but also to control single US if required. The mounting allows for a quick access to the device. Since there will always be situations where control by platoon members becomes difficult, there is also the role of backup operators. They monitor the swarm of US from a workstation in a support vehicle and can take control when the platoon's resources are bound (e.g. in a firefight). Due to their more protected environment, their UI can be

more complex and thus allow for more fine-grained controls. Their workstation also functions as an interface to higher command structures. This is analogous to attached reconnaissance units, who normally bring the US equipment and a lightly armored scout vehicle with them.

3.3 Hardware Used

Two Garm 2 UGVs (Fig. 1 left) and two AirRobot 200 UAVs (Fig. 1 middle) were used. For more information on the robotic systems, see e.g. Brüggemann et al. [2]. The tablet used was a Nexus 9 (HTC). It measures $228.25 \times 153.68 \times 7.95$ mm and weighs 436 g. Its display measures 226 mm diagonally with a resolution of 2048×1536 px. The backup operator's vehicle workstation had a notebook with a 17" 1920×1080 px display with mouse and keyboard input. The squad leaders used Nexus 5 smartphones (LG) in custom-built forearm-mounts (see Fig. 1 right). The Nexus 5 measures $137.8 \times 69.1 \times 8.6$ mm and weighs 130 g. Its display measures 126 mm with a resolution of 1920×1080 px.



Fig. 1. Garm 2 (left), AirRobot 200 (middle) and forearm-mounted Nexus 5 (right).

3.4 UI-Design and Implementation

According to Young and Kott there are roughly three primary paradigms for multi-robot control: Switching, playbook or policy-based control [35]. Switching assigns individual tasks and adjusts coordination if needed. Playbook uses pre-set plans with appropriate parameters. Policy-based control is rule-based and uses dependencies between actors to react to given situations. Our UI is best described as a hybrid between playbook and policy-based control. It consists of a map, a list of orders and units each, a unit detail view and an event ticker (Fig. 2). The map displays own units, US' waypoints and spotted objects. The detail view is available for own and spotted units. In the latter case, time of detection, underlying sensor data (e.g. imagery) and the possibility for a quick classification of the object's type, strength and behavior is provided. Orders can have different connotations in the form of a neutral behavior, as well as aggressive and defensive behaviors. Examples are following a detected object (aggressive), ignoring it (neutral) or avoiding it (defensive). Parameters for different orders are collected by a wizard. Default parameters are selected initially, to allow for skipping details. The system features implicit interaction, e.g. formation keeping and

automatic task allocation. While this influences the users' possibilities for interaction, it is not part of this work for being more on the vehicle side of interaction. The base UI was born out of a testing environment for interoperability. As a custom tool for use by specialists, it was deemed unfit for the actual use case and optimized in this regard. The context of use includes environmental, mission and device constraints. Other infantry activities and the different displays had to be considered. In the following, the UI components are presented. The UI's core is an interchangeable map. As of now, a topographic and a satellite map, switchable by button, are implemented. This function can be extended to include more map types and also map overlay functions.

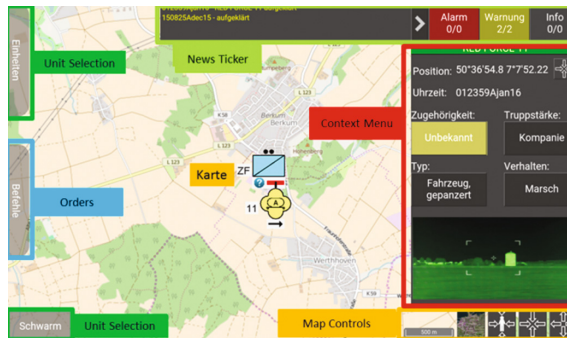


Fig. 2. The different components of the user interface.

The controls are mostly analogous to commercial map applications, like pinch and zoom. Additional buttons are added for centering on (1) the user, (2) the selected unit(s) and (3) on the whole unit. For points (2) and (3) there can be multiple units selected at once, in which case the map zooms to the bounding box containing all selected units. The surrounding area and thus potentially relevant contacts are kept in view. The change of the viewport is animated to ease spatial orientation. The zoom scale is displayed near the map controls. Units are symbolized by according to NATO APP6. As a fast way to assess US status, we implemented a binary color coded bar (red/green) representing general state and an icon for the current task. Units displayed in the map are selectable to access details. The exact coordinates and further information depending on the type of unit is shown in a pop-up panel on the right display border (Fig. 3 left). For own US, more details on status and current orders are displayed. Image sensor readings are displayed as a thumbnail which is switchable to full screen. When the detail view is open, the current sensor orientation is represented on the map. Showing it only on request avoids cluttering the map for larger swarm sizes. Reconnoitered units can be classified by affiliation, strength, type and behavior (Fig. 3 right).

The choices are strongly simplified to avoid overburdening the user. The tab "orders" on the left display border brings up a list of orders reflecting the capabilities of the selected units, e.g. a system without image sensors would not offer image capturing (Fig. 4 left). The selected units are permanently displayed in the lower left display corner and can be changed through the "units" tab. Issuing orders is implemented as a

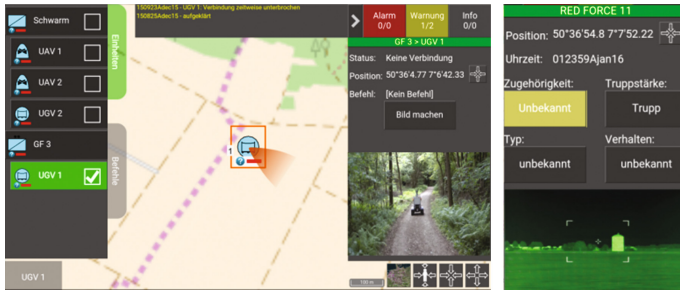


Fig. 3. Context menu with detail information on own system and sensor orientation (left); Unknown object with time, position and the quick classification (right). (Color figure online)

wizard, which leads the user through the required inputs. Orders without parameters, e.g. “Cancel order”, are issued directly. Depending on the type of parameters, there are different input possibilities. The overall behavior of the US can be set through the buttons offensive – neutral – defensive, neutral being the default (Fig. 4 right). The current selection is highlighted. Geographical coordinates can be chosen directly through the map. When editing single coordinates, they can be reset by drag and drop or by just setting another point (Fig. 4 right). For multiple coordinates only drag and drop works. Coordinates can also be deleted through a pop-up context menu. General settings are realized by simple buttons. Pre-set choices are realized like this, e.g. choosing a target squad leader from a list for controlling an US. After issuing an order, the user is notified by a so-called toast as a confirmation.

It is temporarily shown in the lower display center (Fig. 4 left). As mentioned before, units can be selected through the “units” tab on the left display border (e.g. Fig. 3). All entries fulfill two functions: Units can be selected and deselected on the right side of each entry and the left side opens up the information panel for the details on this unit. The currently selected units are always displayed on the lower left. They are furthermore highlighted with an orange bounding box in the map view. Different types of information are generated in the overall system and they have different priorities for the user. This is represented in the UI by three different categories: Info, warning and alert (e.g. Fig. 3). Info contains all notifications with informative character, e.g. feedback from an US about reaching a target waypoint or that it is now controlled by a squad leader. The warning category contains mainly technical problems. These could e.g. be a lost radio connection or a stuck UGV. The alarm category finally contains all threats for the own unit. Examples are detected unknown or enemy contacts in the closer vicinity, sniper detection and NRBC-sensor alarms.

Notifications are displayed in the color code of their categories and a military time stamp in a news ticker at the upper border of the display. Info is coded in white, warnings in yellow and alarms in red. Notifications are also collected regarding their category. Each category can be accessed through respective buttons to the right of the ticker to access a list filtered by the category. The text on these buttons additionally indicates the number of read and unread messages. In the filter views, a list containing the notifications sorted by date is opened at the right border of the display.

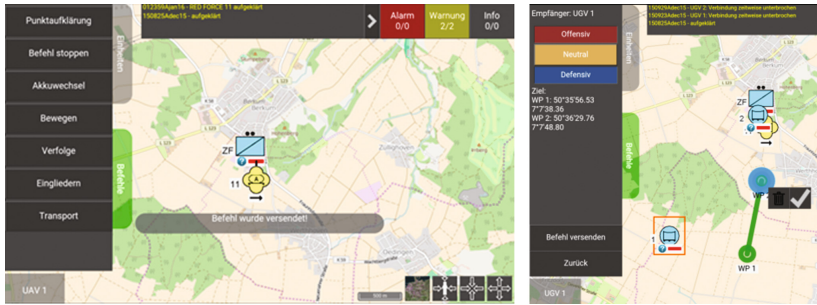


Fig. 4. Units and command sent feedback (left); behavior selection and waypoint input (right).

Each notification represents a link to the coordinates associated with it, which is either the reporting unit or the coordinate of a detected event. Unread notifications are displayed in bold print. Notifications which are no longer needed can be deleted through an additional button. Activating the category button again closes the filter view. Apart from running on different devices, the UI is also adapted software-side to the respective roles to better fit the different contexts of use. These adaptations are presented in the following. The leader UXS is intended to wield the primary control over the overall system of US. The UI was presented in depth above and only serves as a reference for the other roles here. The squad leaders are mainly information consumers in this system. They can access a minimalized UI on their smaller devices. The news ticker is smaller in size and thus presents fewer notifications at once.

The main information medium is to be the digital map with the current situational information. The context menu can only be shown alternatively to the map, since the display size is too small to show both at once with acceptable text and button sizes, which would be preferable [36]. In this context menu, the display is split into a video-feed and the detail view itself. Should a squad leader be given command over an US, the UI generally offers the same functionality, albeit with reduced display size which will affect performance in terms of time [37]. The monitoring and supporting function of the operator requires more details and permanent video-feeds (Fig. 5). The vehicle working place offers more space and protection, so this is not as large a problem as for the dismounted users. The operator has the possibility to directly control US, which is for now implemented for the ground systems. In direct control mode, the respective video-feed is enlarged and a software directional pad is displayed alongside a slider for the speed of the vehicle (Fig. 5 right). Stemming from the functional and technical requirements, there is a range of necessary interaction possibilities for the users. While the displays were described above, this section covers giving orders and the underlying semantics of it. The order “Battery Change” is intended for the flying systems as of now. It serves to order a system to return to the platoon and land to enable a battery change before an automatic return on critical battery state is activated. The order requires a UAV to be selected with no additional parameter.

To cancel orders which were given in error or which are just not needed anymore, the “Cancel Order”-order can be given. Addressed units will stop their current order. No further parameters are needed. The “Move” order requires a selected US and two

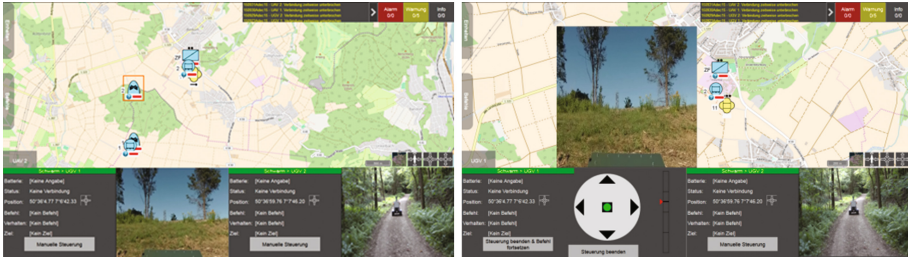


Fig. 5. User interface for the operator role. With enlarged video-feed on the right.

parameters. The general behavior is set to neutral by default, but can also be defensive and offensive. The second parameter is a coordinate on the map which has to be selected by tapping on it. The button “Send Order” finally submits the order to the system. Another way to move US is to let them “Follow” something else. As of now, this is restricted to ground systems and following specific squad leaders. The squad leaders can be selected from a list after which the order can be sent. Apart from automatic reconnaissance while on the move, US equipped with image sensors can be ordered to “Take a Photo” of a specific coordinate. Selecting the coordinate is analogous to the “Move”-order, but the selected point is the photo-target.

The photo will be taken when the system is in sensor range of the target. The leader UXS can “Assign US to Squad Leader” if the need arises. As of now, the order can only be executed for single systems and the only parameter required is the target squad leader, which is chosen from a list analogous to the “Follow”-order. If the platoon stops for longer durations, the swarm of US can be ordered to go into a “Safeguarding” formation around the halting point. The order does not require any parameters. For a “Transport”-order there are two waypoints needed, which are then automatically patrolled by the US, including pre-set halts to load or unload cargo. Setting the waypoints is analogous to the “Move”-order, but additionally a once-set point can be moved by drag-and-drop activated by a long press. Alternatively, points can be deleted by a popup context menu, accessible with a single tap on the point.

Units assigned to squad leaders can “Return to Formation” by the respective order. The loss of control is displayed to the user in question. Apart from a selected US which is currently assigned to someone else, no parameters are needed. Getting into “Formation” means a marching formation in this case. The order is initially sent to all US, which then accordingly arrange themselves in regards to the human platoon members. All current orders are overwritten by this order. It can be given only to the whole swarm. The “Manual Control” on the mobile devices are a backup control for line-of-sight operation. They are designed analogous to the operator, providing a software directional pad and a speed bar. Due to the necessary awareness for the controlled US, the rest of the UI is not available, except for alarm notifications. After the last functionalities were implemented, a final size adjustment of text and interaction elements for the respective devices was performed. Too small elements negatively affect the errorless usage [38, 39], which is especially problematic in combination with the parameterless and thus directly submitted orders.

4 Instruction

The introduction into the controls of the multi-UXV system follows the argumentation of Vogel-Walcutt et al. [40] regarding instructional strategies for military training. Without direct instruction, especially novices to the respective application field tend to turn to inefficient problem solving strategies by using their limited knowledge in largely random search or trial-and-error strategies respectively [41]. Direct instruction is therefore particularly, but not exclusively, a necessary part of optimally designed training processes for novices [42, 43]. With growing expertise, training has to be adapted to avoid redundant or unnecessary instructions [44]. Vogel-Walcutt et al. [40] characterize instructional strategies by the phase of the training cycle (Pre-, During-, Post-training), the experience level of the trainee (novice, journeyman, expert) and the type of knowledge to be conveyed (declarative, procedural, conceptual and integrated). Knowledge typically consists of facts, procedures, concepts, strategies and beliefs [45, 46]. Based on this, Vogel-Walcutt et al. [40] differentiate between declarative, procedural, conceptual and integrated knowledge. Declarative knowledge is base fact knowledge, procedural knowledge entails the knowledge of work steps for task completion, conceptual knowledge is knowledge about relations between information fragments and integrated knowledge is knowledge which is fused with existing knowledge and can be applied to new situations. In short, knowledge types range from solitary knowledge fragments to their linking relations in an application context. While declarative knowledge can be memorized with relative ease, integrated knowledge requires a deeper understanding of the domain. Regarding the re-orientation of the German army's education and training [47], acquiring integrated knowledge can be interpreted as the main goal of educational processes.

Only after the realization of how to integrate new knowledge into a field of duty and own experience it can be expected that training content will be applied even under unexpected circumstances. Minimally directed instructions, e.g. explorative or problem-oriented learning, expect the trainee to find the underlying principles of the training content him- or herself. More directed approaches can however better accommodate to cognitive capabilities of the trainee, which was shown in controlled and randomized experiments [43, 48]. Only after building a base of domain knowledge can the strengths of less directed approaches be leveraged [49]. Since knowledge and competency acquisition are gradual processes, instructions should accordingly be adapted to the progress [50]. The personnel to be trained were available only for a short time, so an adaption was not performed. Military operations are performed by teams of soldiers, making coordinated skills and effective communications as important as individual aspects [51]. Since the UXS swarm is integrated into existing structures, only the coordinated usage of the swarm is of concern. The pre-training phase prepares the training and addresses mainly novice level trainees, which are expected in our case. Information material about the project goals and their role in using the system were provided beforehand. This helps building an initial seed for integrative knowledge for their field of duty.

Right before the introduction began, the soldiers were told about the goals of the introduction to ease building conceptual knowledge. The during-training phase consisted

of presentation, instruction and exercise of the training content. The content was delivered through the scientific personnel, pocket-cards and the system itself. This parallel multimedia presentation can further the intake of procedural, conceptual and integrated knowledge. Usage of the system is trained in three tasks of rising complexity. This segments the overall process and helps understanding. Training with the real system builds procedural knowledge. The training tasks are solved under supervision, providing immediate feedback. Furthermore, a sample solution is generated and presented during this process. Error-correction entails the background of the error, which eases conceptual understanding and following this, integration into existing knowledge. A supervised training task is preferred for novices, since the base knowledge for open problem solving is not yet acquired. The training was performed on military exercise grounds and thus a known environment for the soldiers, which can aid in the transfer to the own working environment. The post-training phase forgoes testing the learned knowledge and consists mainly of a summary feedback of the training. Background information for success and failures are given for further reflection on the learned content and to aid conceptual and integrated knowledge acquisition. The goal of the instruction was to enable the soldiers to operate the UXS-system safely and to make them aware of special situations. The instruction took about two hours. Its four main aspects are presented in the following.

The soldiers were given details on the US and the communication features to better understand the equipment. This included technical and scenario-wise limitations. They were provided with information on UAV flying height, vehicle speeds, sensor loads, automatic detection, ranges, battery change procedure and safety measures during the scenario run. The leader UXV received an introduction into his role as the coordinator of the US and his interfacing role for the platoon leader. He had the opportunity to make himself acquainted with the tablet and the hand carrying loop. The remaining squad leaders received information on their role as mainly information consumers and also had the opportunity to try out the mounting of the devices. Analogous to the set-up of the pocket-cards, the elements of the map, unit selection and commands, context information and the news-ticker were explained (Fig. 6).



Fig. 6. User interface elements explained on the pocket-card (in German).

The scenarios were discussed with the platoon leader. An introduction to the scenario and the plan how to use the US' capabilities were performed by him. As such we could observe how the system would be used when no further constraints are in place.

5 Evaluation

The project presentation centered on a demonstration of the overall system capabilities. German mountain infantry participated in evaluating the UI. They were observed while using the system. Their remarks were protocolled and questionnaires were filled out by them. A realistic scenario was provided to ease the transfer of results to reality.

5.1 Scenario Implementation

Several scenarios were planned beforehand. With the feedback from the platoon leader and the procurement bureau, two scenarios were chosen. In scenario 1 "Reconnaissance on the march", the platoon is operating in known terrain and current maps are existent. The threat level is unknown, but scattered enemy forces are to be expected. The platoon is to redeploy towards a 2 km distant position. The existent road network is to be used. The platoon moves on the road towards the assigned position. No enemy activity is implemented in this scenario. It served to show that the US can follow the platoon without slowing it down and that the leader UXS is able to control and monitor the swarm without a significant increase in task load. In scenario 2 "Stationary Control, dynamic support", the situation is analogous to scenario 1. The platoon is split into an observation post and a scout team, which is to push forward ca. 600 m into a small forest. The observation post is to deploy stationary ground sensors.

The UAVs use a rolling deployment pattern to ensure continuous observation from above. To reconnoiter their movement corridor, the scout team squad leader can request a UAV. Enemy small arms fire is coming from the north. An unmanned resupply transport for the scout team is deployed from the troop carrier halting spot once they are in their assigned position. No enemy forces are expected to be on this route. The scout team unloads the cargo at a position behind their post. The scenario served to show that the US are able to detect, localize and report threats by themselves to enable decision makers to react quickly and adequately. Additional intelligence can be gathered by sending out single US. The battery change of the UAS can be accomplished with little time and under mission context. US are able to support securing the stationary post. Additional sensors can be added on the fly into the existing swarm. The leader UXS does not suffer added significant load. Assigning US to other leaders is fast and safe. The UGVs are able to perform a transport between halting spot and scout team and reintegrate into the swarm afterwards. An introduction to the complete system including the hardware platforms and their capabilities lasted about two hours and preceded the scenario runs. Afterwards, soldiers were free in using the system to accomplish their goals. They received pocket-cards for reference, but field usage was not observed. During the final exercise an observer from the German Army Concepts and Capabilities Development Center participated. He and the platoon leader provided reports on utility and usability of the overall system.

5.2 Short Evaluation

A short usability test was performed to check the UI concept's feasibility. Users had to solve representative tasks with the system. To observe the intuitive use, the leader UXS had to solve three short tasks before the introduction to the system. He was told that the tasks were no means to check his skills, but to find weaknesses in the system. The situation for the tasks was being a leader of four US. The specific tasks were to: (1) Reconnoiter a chosen coordinate with UAV 1 in offensive mode. (2) Open the camera view of UGV 1. (3) Classify an object as an entrenched enemy armored vehicle. The observation yielded insights for different UI-aspects. Users primarily searched the map view to find units. Drag and drop was tried out to move units. The function "Get Picture" was unclear. Feedback by the news ticker was perceived to be helpful. Regarding the context menu, a function to turn the camera, a file administration for a reconnaissance results overview and a send function for imagery were missed. Additional information was missed regarding movement direction, sensor orientation and battery status. Finally, confirming waypoint input was unclear.

5.3 Questionnaires

After the exercises, $n = 6$ soldiers participated in two short questionnaires, the intuitive interaction (INTUI) [52] and the System Usability Scale (SUS) [53], both in German. The INTUI is an instrument to capture intuitive use. The users' replies are depicted in Table 1 and Fig. 7. The highest value is 7; the lowest value is 1. With opposing statements, perceived interaction with a product is reported for four components. A high value in Effortlessness (E) indicates a perceived effortless interaction and low attentional requirements. This matches classical usability the most. For Gut Feeling (G) a high value indicates more intuitive than rational interaction, which is an important feature of intuitive decision making in decision psychology and also in everyday language. A high value for Magical Experience (X) indicates a high quality experience. The interaction is perceived as extraordinary and fascinating. It has a broader scope than task fulfillment alone, analogous to user experience (UX).

For Verbalizability (V) high values state that interaction over time is well describable, which indicates a logical sequence of interaction steps. However, the origin of relevant knowledge for intuitive use is often unconscious and thus not verbalizable.

Table 1. Descriptive statistics of the of the INTUI results.

Scale	M	SD
Effortlessness (E)	5,3	0,575
Gut feeling (G)	3	1,089
Verbalizability (V)	5,9	1,068
Magical experience (X)	5	1,072

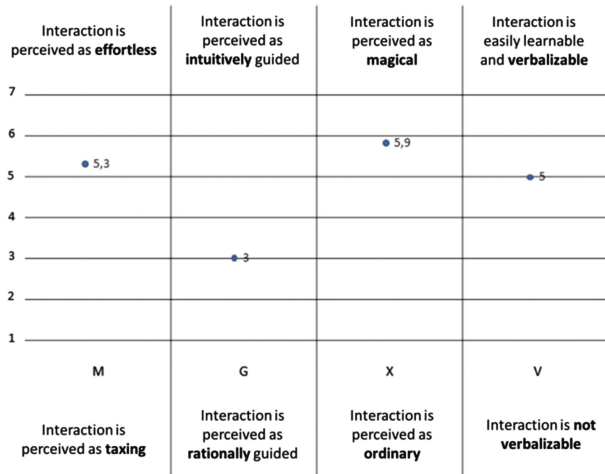


Fig. 7. Graphical depiction of the INTUI results.

Interaction consists of many small, unconscious decisions, which together form a meaningful, verbalizable sequence. The System Usability Scale questionnaire measures experienced usability as the so-called SUS-score and consists of ten items (see Fig. 8). It ranges from 0 (worst application imaginable) to 100 (best application imaginable). It is used to determine perceived user friendliness of software. The questionnaire contains five positive and negative statements each, regarding the usability of the system. The UI scored 76.25, which equates to good usability.

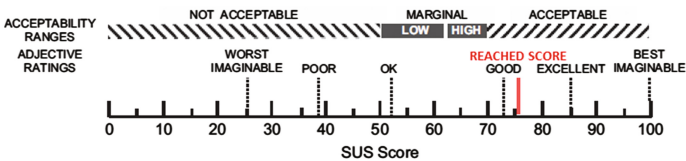


Fig. 8. The SUS-score range.

5.4 Field Observations and Comments

Observations and user comments from the evaluation were protocolled. Handedness was discussed. Keeping the dominant hand mostly free will obviously increase acceptance. The leader UXS wished for something similar to the forearm-mounts. Providing mountings and storage suited for both left- and right-handers are a general requirement for system introduction. Mounting a tablet-sized device to the forearm however restricts flexibility, so there is no trivial solution to this problem. At first, the vanguard UGV was relatively close to the platoon (75 m). The soldiers reported being drawn to march faster, possibly leading to security risks. A larger distance was perceived to be much better, but it remains an open question, to which extent the soldiers

are still influenced. Orders for sub-groups of the US were requested by the soldiers. This was discussed but not implemented due to the small swarm size. UGVs were seen as a platform for transport, IED-detection, jammers and sniper detection. Soldiers wished for alarms to be transmitted when a soldier is not looking at the device, which can possibly be implemented by vibro-tactile or audio cues.

The visualization of the sniper detection was found to be unintuitive. Shot direction and source as well as how much shooting occurred (caliber, cadence, etc.) were difficult to determine. Improved filters for own fire were requested. Wishes regarding display duration for shot detection differed between ranks: For the platoon leader five minutes and more were discussed, while for squad leaders 90 s should suffice. It was discussed if the displayed image of US should automatically switch to movement direction and if a scrollable 360° feed could help. Despite the need to detach the swarm from the platoon leader by introducing a proxy role, the platoon leader noted that quick access to sensor visualizations would be helpful for him. Grid visualization with thumbnails of available sensor readings could be a solution which also helps squad leaders without direct command of US with faster access to the sensors. Also, remaining battery charge or alternatively the estimated remaining flight time was requested by the soldiers. Manual input of observations was requested to add to the situational overview. The possibility to manually mark objects in video feeds was requested again, after being mentioned during the pre-implementation interviews, to compensate failures in automation. The size of the smartphones and the resulting decision to only show the video feed itself lead to mapping problems between feed and observed map section. The idea to integrate traditional radio communication for the interaction between consumers and controllers was generally received well, but deeper integration was requested to eliminate device switching. Finally, a planning mode for pre-recording action sequences was requested.

6 Conclusion and Future Work

The project was intended to make a step forward towards the introduction of US into the German armed forces on a swarm scale. The success of such an undertaking is not only a technical problem, which is why the project had three goals: (1) Study of the technical challenges, like the necessary level of automation in control and sensor data processing, but also the interoperability, technical communication and architecture. (2) Study of the integration of US into an infantry platoon. This means integrating into existing structures and processes to create a soldier-multi-robot-team in which the US can be precisely controlled even in stressful situations. (3) Proof of concept with a real test run. The diverse methods and requirements, many of whom were generated together with intended users of the system, needed to be integrated into a whole. Only with such an overarching system all concepts and ideas can be evaluated and checked for their benefits and disadvantages. A real demonstrator also shows how long the way from research to the real use of the system still is. Progress was made in all areas. The UI implementation was well received by the participating soldiers. However, there are still many aspects to be improved. An aspect which came up during the interviews was using different in- and output modalities, which were restricted to visual output and

touch input in this project. Audio and haptics could be used as alternative output modalities, as long as the notifications are discreet but noticeable enough. This also addresses notifying users who are not looking at the display.

A study by Oron-Gilad et al. reports benefits of additional tactile cues in simulated hostile environments [54]. Alternative visual displays could be new data glasses or more generally Helmet Mounted Displays (HMDs). Further possible input modalities can be gesture, speech, head and/or eye tracking. Some work in this area exists already and will function as a basis [15, 55]. The very concrete scenario of the tests leaves open the questions how generalizable the elements of the UI are or if elements are missing. Mission planning functions will help to accommodate different use cases. Ophir-Arbelle et al. studied how video feeds from UGV and UAV are used by dismounted soldiers [56, 57]. They could show that combined ground and air footage improves the reconnaissance performance. Such combinations of US are possible in the system we used, so the use of a combined view is worth looking into. Another important aspect is system security. Jeong and Ha present a design and implementation of a secure architecture for multi-user situations in UAV control [58]. Security considerations can impact the UI design, especially active mechanisms like rights management. We plan to address open aspects in future projects and to further develop the possibilities US can provide to dismounted soldiers in their dangerous tasks.

References

1. Stimpert, S.S.: Lightening the load of a USMC Rifle Platoon through robotics integration. DTIC Document, January 2014
2. Brüggemann, B., et al.: Manned-unmanned teaming to support a dismounted infantry platoon. In: IST-127/ RSM-3, Intelligence and Autonomy in Robotics (2016)
3. Remmersmann, T., et al.: Towards duty - BML communication enables a multi-robot system supporting an infantry platoon. In: Proceedings of the SISO 2016 Fall Simulation Innovation Workshop (2016)
4. Sterling, B.S., Perala, C.H.: Workload, stress, and situation awareness of soldiers who are controlling unmanned vehicles in future urban operations. DTIC Document (2007)
5. Sterling, B.S., Perala, C.H.: Controlling unmanned systems in a simulated counter-insurgency environment. DTIC Document (2007)
6. Redden, E.S., Elliott, L.R., Pettitt, R.A., Carstens, C.B.: Scaling robotic systems for dismounted warfighters. *J. Cogn. Eng. Decis. Making* **5**(2), 156–185 (2011)
7. Langerwisch, M., et al.: Realization of an autonomous team of unmanned ground and aerial vehicles. In: International Conference on Intelligent Robotics and Applications, pp. 302–312 (2012)
8. Remmersmann, T., Schade, U., Schlick, C.M.: Interactive multi-robot command and control with quasi-natural command language. In: 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 470–475 (2014)
9. Remmersmann, T., Schade, U., Tiderko, A.: Commanding heterogeneous multi-robot teams. DTIC Document, January 2014
10. Mi, Z.-Q., Yang, Y.: Human-robot interaction in UVs swarming: a survey. *Int. J. Comput. Sci. Issues* **10**(2), 273–280 (2013)

11. Prinet, J.C., Terhune, A., Sarter, N.B.: Supporting dynamic re-planning in multiple UAV control: a comparison of 3 levels of automation. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 423–427 (2012)
12. Morfopoulos, A., McHenry, M., Matthies, L.: Reduction of user interaction by autonomy. DTIC Document, January 2004
13. Barnes, M.J., et al.: Designing for humans in autonomous systems: military applications. DTIC Document, January 2014
14. Chen, J.Y.C., Barnes, M.J.: Human-agent teaming for multirobot control: a review of human factors issues. *IEEE Trans. Hum.-Mach. Syst.* **44**(1), 13–29 (2014)
15. 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**(6), 1231–1245 (2007)
16. Barnes, M.J., Chen, J.Y.C., Jentsch, F.: Designing for mixed-initiative interactions between human and autonomous systems in complex environments. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1386–1390 (2015)
17. Stroumtsos, N., Gilbreath, G., Przybylski, S.: An intuitive graphical user interface for small UAS. In: SPIE Defense, Security, and Sensing, p. 87410F (2013)
18. Mercado, J.E., et al.: Intelligent agent transparency in human-agent teaming for multi-UxV management. *Hum. Factors* **58**(3), 401–415 (2016)
19. Clare, A.S., Ryan, J.C., Jackson, K.F., Cummings, M.L.: Innovative systems for human supervisory control of unmanned vehicles. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 531–535 (2012)
20. Clare, A.S., Macbeth, J.C., Cummings, M.L.: Mixed-initiative strategies for real-time scheduling of multiple unmanned vehicles. In: American Control Conference (ACC 2012), pp. 676–682 (2012)
21. Franchi, A., Secchi, C., Ryll, M., Bulthoff, H.H., Giordano, P.R.: Shared control: Balancing autonomy and human assistance with a group of quadrotor UAVs. *IEEE Robot. Autom. Mag.* **19**(3), 57–68 (2012)
22. Cook, M.B., Smallman, H.S., Lacson, F.C., Manes, D.I.: Guided attention for autonomous system supervision. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 408–412 (2012)
23. Taylor, G., Frederiksen, R., Crossman, J., Quist, M., Theisen, P.: A multi-modal intelligent user interface for supervisory control of unmanned platforms. In: 2012 International Conference on Collaboration Technologies and Systems (CTS), pp. 117–124 (2012)
24. Endsley, M.R.: From here to autonomy: lessons learned from human-automation research. *Hum. Factors* **59**, 001872081668135 (2016)
25. Dawson, S., Crawford, C., Dillon, E., Anderson, M.: Examining the expectations of autonomy and human intervention in a multi-robot surveillance task. In: Proceedings of the 50th Annual Southeast Regional Conference, pp. 345–346 (2012)
26. Demir, M., McNeese, N.J., Cooke, N.J.: Team situation awareness within the context of human-autonomy teaming. *Cogn. Syst. Res.* (2016)
27. Evans III, A.W., Hill, S.G., Pomranky, R.: Investigating the usefulness of soldier aids for autonomous unmanned ground vehicles, Part 2. DTIC Document, January 2015
28. Peschel, J.M., Duncan, B.A., Murphy, R.R.: Exploratory results for a mission specialist interface in micro unmanned aerial systems. In: 2012 International Conference on Collaboration Technologies and Systems (CTS), pp. 131–140 (2012)
29. Pitman, D., Cummings, M.L.: Collaborative exploration with a micro aerial vehicle: a novel interaction method for controlling a mav with a hand-held device. *Adv. Hum.-Comput. Interact.* **2012**, 18 (2012)

30. Solovey, E., Jackson, K., Cummings, M.: Collision avoidance interface for safe piloting of unmanned vehicles using a mobile device. In: Adjunct proceedings of the 25th annual ACM symposium on User interface software and technology, pp. 77–78 (2012)
31. Rice, T.M., Chhabra, T., Keim, E.A.: Unmanned Tactical Autonomous Control and Collaboration Concept of Operations. Naval Postgraduate School, Monterey (2015)
32. Bommer, S.C.: Assessing the Effects of Multi-Modal Communications on Mental Workload During the Supervision of Multiple Unmanned Aerial Vehicles (2013)
33. Mazzara, M.: UGV Interoperability Profile (IOP) Capabilities Plan for Version 0. DTIC Document, January 2011
34. Oron-Gilad, T., Parmet, Y.: Close target reconnaissance a field evaluation of dismounted soldiers utilizing video feed from an unmanned ground vehicle in Patrol Missions. *J. Cogn. Eng. Decis. Making* (2016)
35. Young, S., Kott, A.: A survey of research on control of teams of small robots in military operations. arXiv preprint, [arXiv:1606.01288](https://arxiv.org/abs/1606.01288) (2016)
36. Hou, M., Ho, G., Arrabito, G.R., Young, S., Yin, S.: Effects of display mode and input method for handheld control of micro aerial vehicles for a reconnaissance mission. *IEEE Trans. Hum.-Mach. Syst.* **43**(2), 149–160 (2013)
37. Redden, E.S., Pettitt, R.A., Carstens, C.B., Elliott, L.R., Rudnick, D.: Scaling robotic displays: visual and multimodal options for navigation by dismounted soldiers. DTIC Document, January 2009
38. Conradi, J., Alexander, T.: Analysis of visual performance during the use of mobile devices while walking. In: International Conference on Engineering Psychology and Cognitive Ergonomics, pp. 133–142 (2014)
39. Conradi, J., Busch, O., Alexander, T.: Optimal touch button size for the use of mobile devices while walking. *Procedia Manufact.* **3**, 387–394 (2015)
40. Vogel-Walcutt, J.J., Fiorella, L., Malone, N.: Instructional strategies framework for military training systems. *Comput. Hum. Behav.* **29**(4), 1490–1498 (2013)
41. Sweller, J.: Instructional design. In: Australian Educational Review (1999)
42. Mayer, R.E.: Should there be a three-strikes rule against pure discovery learning? *Am. Psychol.* **59**(1), 14 (2004)
43. Kirschner, P.A., Sweller, J., Clark, R.E.: Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ. Psychol.* **41**(2), 75–86 (2006)
44. Renkl, A., Atkinson, R.K.: Structuring the transition from example study to problem solving in cognitive skill acquisition: a cognitive load perspective. *Educ. Psychol.* **38**(1), 15–22 (2003)
45. Bloom, B.S.: Taxonomy of educational objectives: the classification of educational goals (1956)
46. Krathwohl, D.R.: A revision of Bloom’s taxonomy: An overview. *Theory Pract.* **41**(4), 212–218 (2002)
47. Marstaller, A.: Ausbilden für die Streitkräfte: Kompetenzerwerb in der Bundeswehr, Hamburg, 2 September 2015
48. Plass, J.L., Moreno, R., Brünken, R.: *Cognitive Load Theory*. Cambridge University Press, New York (2010)
49. Kalyuga, S., Ayres, P., Chandler, P., Sweller, J.: The expertise reversal effect. *Educ. Psychol.* **38**(1), 23–31 (2003)
50. Walsh, M.B., Moss, C.M., Johnson, B.G., Holder, D.A., Madura, J.D.: Quantitative impact of a cognitive modeling intelligent tutoring system on student performance in balancing chemical equations. *Chem. Educ.* **7**(6), 379–383 (2002)

51. Salas, E., Cooke, N.J., Rosen, M.A.: On teams, teamwork, and team performance: discoveries and developments. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **50**(3), 540–547 (2008)
52. Ullrich, D., Diefenbach, S.: INTUI. exploring the facets of intuitive interaction. In: *Mensch & Computer 2010*, p. 251 (2010)
53. Brooke, J.: SUS-A quick and dirty usability scale. *Usability Eval. Ind.* **189**(194), 4–7 (1996)
54. Oron-Gilad, T., Parmet, Y., Benor, D.: Interfaces for dismounted soldiers examination of non-perfect visual and tactile alerts in a simulated hostile urban environment. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pp. 145–149 (2015)
55. Lackey, S., Barber, D., Reinerman, L., Badler, N.I., Hudson, I.: Defining next-generation multi-modal communication in human robot interaction. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pp. 461–464 (2011)
56. Ophir-Arbelle, R., Oron-Gilad, T., Borowsky, A., Parmet, Y.: Is more information better?: How dismounted soldiers use video feed from unmanned vehicles: attention allocation and information extraction considerations. *J. Cogn. Eng. Decis. Making* **7**(1), 26–48 (2013)
57. Oron-Gilad, T., Parmet, Y.: Is more information better for dismounted soldiers?: Display-layout considerations of multiple video feed from unmanned vehicles. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **58**(1), 345–349 (2014)
58. Jeong, H.-J., Ha, Y.-G.: Design and implementation of secure control architecture for unmanned aerial vehicles. *Int. J. Smart Home* **7**(3), 385–392 (2013)

“Smooth” or “Intermittent”? The Necessity of Halt in the Dynamic Visualization Due to the Features of Working Memory

Xiaozhou Zhou¹, Chengqi Xue^{1(✉)}, An Li², Yafeng Niu^{1,3},
and Jing Zhang¹

¹ School of Mechanical Engineering,
Southeast University, Nanjing 211189, China
ipd_xcq@seu.edu.cn

² Shanghai Ucloud Info Tech Ltd., Shanghai 200090, China

³ Science and Technology on Electro-optic Control Laboratory,
Luoyang 471023, China

Abstract. With the improvement of computer capability and the development of visualization technology, dynamic visualization could be displayed smoothly. However, the span limitation of human’s working memory is a natural barrier to obtain the massive information in parallel. Here, we designed a simple psychological experiment to validate the necessity of the halt in the dynamic visualization. In this 2×2 between-subjects design test, 21 graduate students participated in the control group and experimental group respectively to compare the influence of the halt in the visualization under the conditions of the simple search task and the complex search task. The eye movement data of number of fixation points and total fixation duration of each single material were recorded to investigate the real-time cognitive load. The results showed that the performance improved significantly when the halt added in the complex search task, along with the real-time workload reduced. However, the performance and cognitive load have no significant change in the simple search task. It demonstrated the need for the appropriate halt in the complex data visualization.

Keywords: Dynamic visualization · Working memory span · Eye movement · Cognitive load

1 Introduction

The notion that working memory is the basis for temporary storage and processing the complex information and it does effect the understanding ability of human was generally accepted ever since the seminal studies proposed by Baddeley [1, 2].

Foundation items: The National Natural Science Foundation of China (No. 71471037, 71271053), Science and Technology on Electro-optic Control Laboratory and National Aerospace Science Foundation of China (No. 20165169017), SAST Foundation of China (SAST No. 2016010), The Scientific Innovation Research of College Graduates in Jiangsu Province (No. KYLX_0104).

© Springer International Publishing AG 2017

D. Harris (Ed.): EPCE 2017, Part II, LNAI 10276, pp. 179–188, 2017.

DOI: 10.1007/978-3-319-58475-1_13

Studies have been completed to establish the high correlation between the working memory and the study ability [3, 4]. Other human's cognitive abilities are much higher than the short-term memory storage, thus the working memory should be seen as a short board and the key factor of human's cognitive ability. Due to this, numerous researchers were interested in the working memory.

The big data visualization nowadays with the characteristics of high dimensions and magnanimity is usually displayed in a dynamic form. Along with the development of computer hardware and software technology, the dynamic visualization could reach the close to real-time and presented smoothly. The smoother of the visualization, the better of it, is that reasonable? Since human is the cognition subject, the unique index for evaluating the visualization should be the human's cognitive performance. Due to the complexity of big data, visual cognition is a typical complex task usually accompanied with high cognitive load. Therefore, the visualization mechanism of large database should follow the characteristics of working memory to acquire better cognitive effects. In the experiments of our previous study [5], we noticed that the recall accuracies of the subjects were significant higher in the condition of temporally shield than the condition of skip gradually and the condition of skip directly without halt. This finding made us to pay attention to the relationship between human's working memory features and the information's visualization presentation.

According to the prior studies, working memory includes both the functions of processing and the storage [6, 3]. As for the process of working memory, one description was task-switching resource-sharing model [7, 8], while another was time-based resource-sharing model [9]. There was no essential difference between these two resource-sharing models. According to these resource-sharing models, processing and storage were competition for the limited working memory resources when we receive the novel information. As time progresses, attention switched from processing to the retrieval, refreshing of decaying memory traces was needed to recall the past items. However, previous studies were focused on the phonological loop subsystem (i.e., reading task) of working memory [10, 11]. And since the on-screen visualization is based on the graphical understanding, it involves more of the visuospatial sketchpad subsystem of working memory, our research was more focused on the visuospatial sketchpad subsystem. And the major goal of this study is to establish whether we need to empty our working memory storage in appropriate time during the displaying of dynamic visualization to let the novel information in.

Eye tracking data have already been used to measure human's mental workload in real-time and accurately [12, 13]. The eye tracking measure is better than the traditional performance measure on reaction time or accuracy for its precise temporal reflection of mental process. It can indicate the time, space and duration data about the participants' fixating on certain stimuli. Studies revealed that total duration time index performed a better predictor than other index like blink duration [14]. So, we choose the concurrent eye movement data total duration time and fixation number and the performance data reaction time to estimate the real-time cognitive load in this study.

2 Experiment

2.1 Subjects and Materials

The theoretical background and the results of the previous studies led to the following experiments to verify the validity of the halt in the dynamic visualization. The experiment was 2×2 between-subjects design with one experimental group and one control group. We designed two kinds of experimental material, as shown in Fig. 1. One was arrayed in dot matrix, while another was in annular, and both of them were common to see in data visualization. In the matrix figure, there was 1 target dot and 329 distraction dots, and 2 target dots with line connected and 90 distraction dots in the annular one. The matrix (labeled Vis. 1) was defined as hard task for more interferences in the figure and it's hard to distinguish the similar target dot in it. And the annular (labeled Vis. 2) was defined as easy task since the distraction dots were much less than the matrix figure and the line connected the two target dots made the subjects identify and judge easier. The target of the searching task was a solid polygon with approximate circular shape while the interfering objects were circular solid points, seen in Table 1. The meaningless of both the target graphic and the distractors could prevent the subjects to use the image thinking during the searching tasks. The subjects were asked to reaction by press the keyboard as soon as they found the target node.

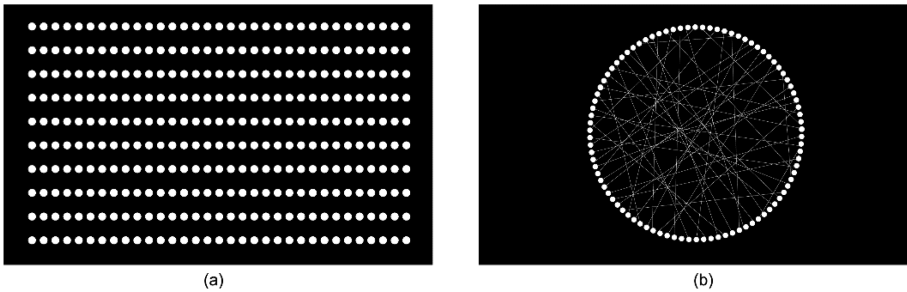




Fig. 1. Two kinds of experimental material (a) was arrayed in dot matrix, the corresponding searching task was hard; (b) was arrayed in annular, the corresponding searching task was easy.

And 21 graduate students participated in each group. The Tobii X2-300 compact contactless eye tracker was used to collect the eye movements' and performance data. The instrument sampling frequency is 30 Hz, the staring accuracy is 0.4° – 0.5° and the head movement range is about 50×36 cm. The resolution ratio of experimental animation material is 1280×960 px. The experimental materials were presented by a HP 21 inch screen with the brightness of 92 cd/m^2 . The laboratory was in the normal lighting conditions (40 W fluorescent). The distance between the subjects and the screen was about 550–600 mm.

Since there was no exact value about the visual short-term memory retention [15, 16], a 700 ms interval of blank screen was setup before each visual graphic in the experimental group based on experience. Both the experimental group and the control group

Table 1. Interference objects and target object of the material in both the experimental and control group

		Interference Node	Target Node	Degree of Complexity
Size		25px	25px	
Sample				
Quantities	Vis. 1	329	1	Hard
	Vis. 2	90	2	Easy

tested two kinds of figure three times (including the trials with homogeneous figure). The only difference between the experimental group and the control group was the 700 ms interval before the figure presented in each trial.

2.2 Procedure

The materials presented in experimental group were as same as in the control group. The procedure of the test trial in each experimental group was showed in Fig. 2. The top row was a test trial in the control group, and the below row was a test trial in the experiment group with a 700 ms blank interval before each test picture. The subjects were tested repeatedly in six homogeneous trials, and in each trail the location of the target point appeared randomly in each test trial while the form of the materials keeping consistent. In order to avoid the effect of the attention blink on the experimental results, a blank picture with “+” was displayed for 500 ms to eliminate the visual residue [17]. The reaction time was recorded from the moment when the picture displayed in the screen to the moment when subjects pressed the keyboard as soon as they found the target.

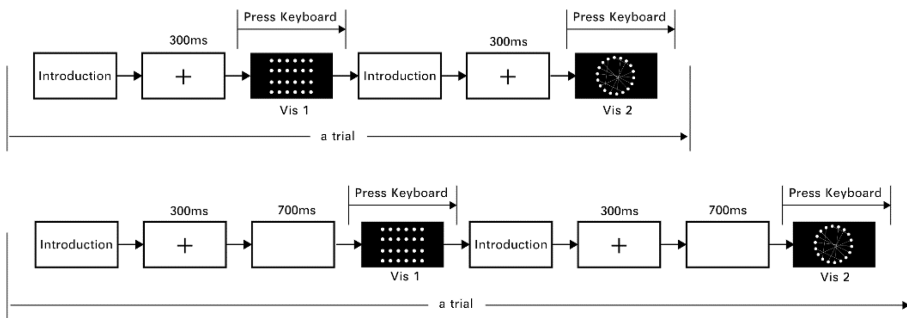


Fig. 2. The procedure of a test trial in each group. The top row is a test trial in the control group, and the below row is a test trial in the experiment group with a 700 ms blank interval before each test picture.

3 Results

The two-factor ANOVA was used to analyze the results with the two factors of task difficulties and blank interval. The response time (RT), the number of fixation time (N) and the total fixation duration (FD) were taken as the dependent variables. Among them, searching time data RT represented the searching performance level and the concurrent eye movement data N and FD could reflect the subjects’ real-time workload situation. As to estimate the situation of performance and eye movement under a certain workload, the data of the first trial in each group were not counted in. The results showed that the average values of searching time, number and the total duration of the fixation were relatively higher in the condition of hard task (Vis. 1) and the condition in control group (without blank intervals). As seen in Fig. 3, the lines of reaction of three factors RT, N and FD had the intersecting tendencies with strong consistency. It implied that the influences of the blank interval were different in the conditions of different task difficulties and the lower performance and the higher real-time workload in the situation when the task got hard or the blank interval was omitted.

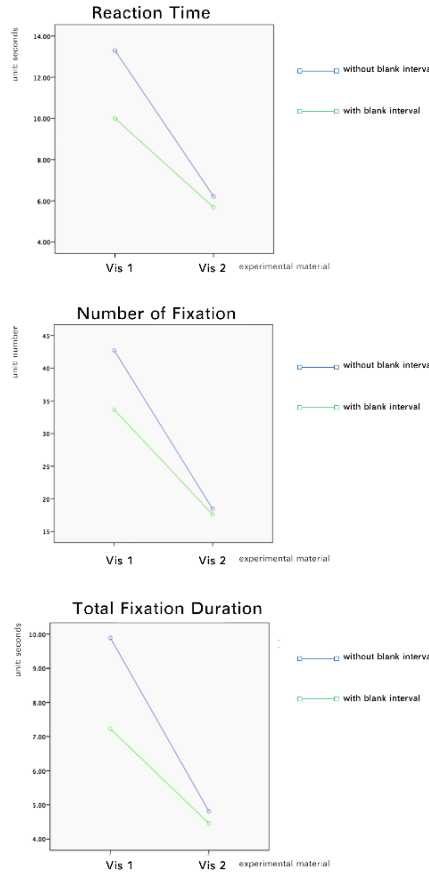


Fig. 3. The interaction appeared in the dependent variables of RT, N and FD.

In particular, seen from the results of pairwise T-test analysis between the control group and the experimental group (Table 2), there were significant differences in RT ($t = 2.163, df = 41, sig. = 0.036 < 0.05$), N ($t = 2.124, df = 41, sig. = 0.040 < 0.05$), and FD ($t = 2.174, df = 41, sig. = 0.036 < 0.05$) in the condition of hard task. And these differences were unidirectional, that were, the performance level increased (RT decreased) when the blank intervals were added. And no significant differences existed in the condition of easy task, with the results of RT ($t = -0.827, df = 41, sig. = 0.413 > 0.05$), N ($t = -1.025, df = 41, sig. = 0.312 > 0.05$), and FD ($t = -0.847, df = 41, sig. = 0.401 > 0.05$).

Table 2. The pairwise sample T-test results between the control group and the experimental group with the dependent variables RT, N and FD

Control group- experimental group		Paired difference					t	df	Sig. (two-sides)
		Mean	Standard deviation	Standard error of the mean	The 95% confidence interval for the difference				
					Lower limit	Upper limit			
Pair 1	RT of Vis1	5.341	16.000	2.469	.355	10.327	2.163	41	.036*
Pair 2	RT of Vis 2	-.642	5.030	.776	-2.209	.925	-.827	41	.413
Pair 3	N of Vis 1	16.643	50.784	7.836	.818	32.468	2.124	41	.040*
Pair 4	N of Vis 2	-2.262	14.307	2.208	-6.720	2.196	-1.025	41	.312
Pair 5	FD of Vis 1	4.042	12.048	1.859	.288	7.797	2.174	41	.036*
Pair 6	FD of Vis 2	-.569	4.343	.670	-1.922	.785	-.849	41	.401

* The mean difference is significant at the 0.05 level (two-sides)

Besides the value of concurrent eye movement data fixation number and total fixation duration, the difference of real-time workload in different tasks could be seen clearly in the heat map, seen in Fig. 4. The top row (a, b) was the heat maps of the saccadic data of a typical subject in experimental group (with halt); the bottom row (c, d) was the heat maps of the saccadic data of a typical subject in control group (without halt). Figure 5 showed the overlapped heat map of total 21 subjects in each group. As same as Fig. 4, the left two maps (a, c) showed the eye-movement heat maps generated by the same experimental figure. We can see the obvious differences in the saccadic number and path between the left two maps (a, c), and these differences hardly be found in the right two maps (b, d). It showed whether there was a halt to be significantly affected the eye movement in the condition of hard task, and no significant difference in eye movement in the condition of easy task, due to the low level of overall workload.

We can see the different searching preferences from the figure of eye movement track. In the trials of matrix figure, subjects always scanned the figure with the habit of reading words. They scanned from left to right and from top to bottom generally, than they wandered around the suspected target to make sure the answer. In the trials of annular figure, most of the subjects browsing along the circle (counterclockwise majority),

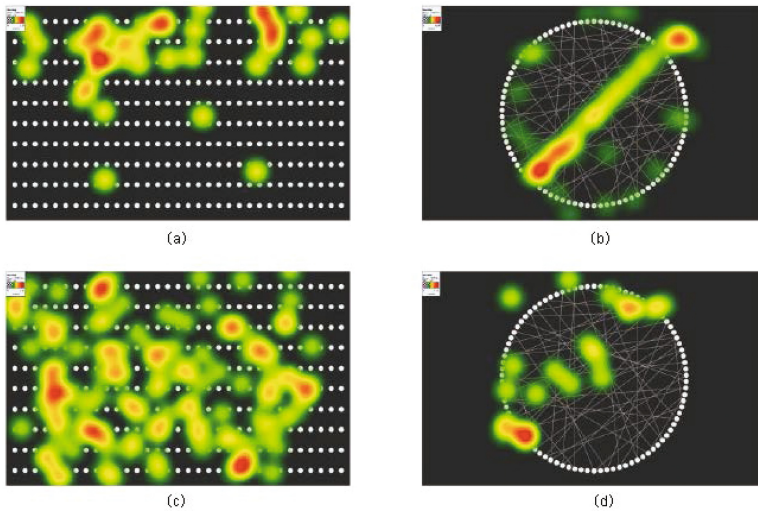


Fig. 4. The heat maps of two selected subjects of the experimental group and the control group. The top row (a,b) is the heat maps of the saccadic data of a subject in experimental group (with halt); the bottom row (c,d) is the heat maps of the saccadic data of a subject in control group (without halt).

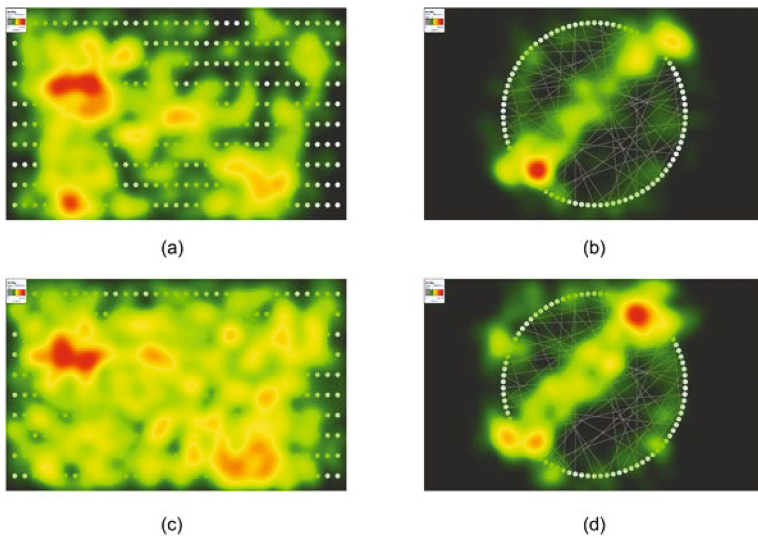


Fig. 5. The overlapped heat maps of 21 subjects of the experimental group and the control group. The top row (a,b) is the overlapped heat maps of the saccadic data of 21 subjects in experimental group (with halt); the bottom row (c,d) is the overlapped heat maps of the saccadic data of 21 subjects in control group (without halt).

than stayed around the suspected targets to determine. In Fig. 5(c), it was hard to see the underlying experimental figure since the gaze points almost distributed on the entire figure. The circular motion of the eyeball was much easier than the repeat horizontal movement. It was one of the reasons that mean reaction time in the annular figures was shorter than in the matrix figures.

In addition, all the *Pearson* correlation coefficients between any two factors of RT, N and FD in each trial were greater than 0.90 and reached the significance at the 0.01 level. This result strongly proved the high degree of correlation between the performance and eye movement. Since the concurrent eye movement index reflected the instantaneous cognitive load, the high correlation between operational performance and cognitive load in the interface task was confirmed again.

4 Discussion

The reason of the results should be attributed to the limitation of working memory span. Based on the previous studies, working memory fulfills the functions of processing and storage at the same time. When the novel information flows into working memory and the amount exceeds the span restriction of the working memory (e.g., in Vis. 1), the processing performance would be undermined. In this study, the processing task was to judge the point to be the target or interference one. And if the tolerance of the novel information were within the span limited of working memory (e.g., in Vis. 2), the processing performance would not be influenced. The experimental results showed that the storage and processing processes competed for limited attention resources when the memory span demands for the current task exceeded what it can tolerate. Since the experimental materials in this study were abstract graphic, which were difficult to convert to verbal memory, the working memory involved in this study was Visuospatial Sketchpad subsystem. And it exhibits the same span limitation as same as the Phonological loop in the previous studies.

Therefore, we can explain the results as the following. When the visual complexity is low, the halt has no significant influence in the cognitive load and the searching performance. While the visual complexity getting higher, the appropriate halt has a positive meaning to reduce the overall workload and improve the search performance. Based on the characteristics of big data visualization, the visual materials which we see in our real-life situation tends to be much more complex than in the experimental conditions (since all the factors of texts, graphics, background, etc. that could increase the overall cognitive load have been removed) in this study. So we can reasonable assume that the dynamic data visualization in reality should be in a complex task condition.

The fact that the interval time before each searching task could reduce the cognitive load and improve the searching performance implied a positive meaning of the halt. So we need not pursue a completely smooth and interval-free dynamic visualization and the appropriate halt is beneficial to the human's understanding. Maybe we need to set a reasonable pause or blank before the query information appears on the path of infinite pursuit of fluid visualization. We should halt for a while and consider people's ability to accept. As we know, visualization is the bridge between the human and the data. The best visualization of data should be the most suitable one for human to read and understand.

In this paper, the definition of complex and simple task was too simple, we need to further define the task complexity to obtain more accurate conclusions. However, this research makes sense with the dynamic visualization design to some extent. Follow-up research of this study is how to set the halt in a visualization to make it not abrupt but friendly.

5 Conclusion

The spatial organization form of visualization objects could affect subjects' visual search path. The nodes arranged in annular would result in higher search performance than in progressive sort. Due to the span limitation and the competition between storage and processing in human's working memory, in the complex cognitive task such as reading big data visualization which inevitably with a high cognitive load, the appropriate pauses and blank would be beneficial to cognition in the course of displaying dynamic visualization. And in condition of simple cognitive task, the halt in the dynamic display is not necessary.

Acknowledgement. This paper is supported by National Natural Science Foundation of China (No. 71471037, 71271053) and Science and Technology on Electro-optic Control Laboratory and National Aerospace Science Foundation of China (No. 20165169017), SAST Foundation of China (SAST No. 2016010) and The Scientific Innovation Research of College Graduates in Jiangsu Province (No. KYLX_0104).

References

1. Turner, M.L., Engle, R.W.: Is working memory capacity task dependent? *J. Mem. Lang.* **28** (2), 127–154 (1989)
2. Baddeley, A.: Working memory and language: an overview. *J. Commun. Disord.* **36**(3), 189–208 (2003)
3. Daneman, M., Carpenter, P.A.: Individual differences in working memory and reading. *J. Verbal Learn. Verbal Behav.* **19**(4), 450–466 (1980)
4. King, J., Just, M.A.: Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* **30**(5), 580–602 (1991)
5. Zhou, X., Xue, C., Zhou, L., Shao, J., Shen, Z.: Spatial conformity research of temporal order information presentation in visualization design. In: Yamamoto, S. (ed.) *HIMI 2016*. LNCS, vol. 9734, pp. 91–99. Springer, Cham (2016). doi:[10.1007/978-3-319-40349-6_10](https://doi.org/10.1007/978-3-319-40349-6_10)
6. Case, R., Kurland, D.M., Goldberg, J.: Operational efficiency and the growth of short-term memory span. *J. Exp. Child Psychol.* **33**(3), 386–404 (1982)
7. Towse, J.N., Hitch, G.J.: Is there a relationship between task demand and storage space in tests of working memory capacity? *Q. J. Exp. Psychol.* **48**(1), 108–124 (1995)
8. Barrouillet, P., Camos, V.: Developmental increase in working memory span: resource sharing or temporal decay? *J. Mem. Lang.* **45**(1), 1–20 (2001)
9. Barrouillet, P., Bernardin, S., Camos, V.: Time constraints and resource sharing in adults' working memory spans. *J. Exp. Psychol. Gen.* **133**(1), 83 (2004)

10. Swanson, H.L.: Reading comprehension and working memory in learning-disabled readers: Is the phonological loop more important than the executive system? *J. Exp. Child Psychol.* **72**(1), 1–31 (1999)
11. Duff, S.C., Logie, R.H.: Processing and storage in working memory span. *Q. J. Exp. Psychol. Sect. A* **54**(1), 31–48 (2001)
12. Desroches, A.S., Joanisse, M.F., Robertson, E.K.: Specific phonological impairments in dyslexia revealed by eyetracking. *Cognition* **100**(3), B32–B42 (2006)
13. Ahlstrom, U., Friedman-Berg, F.J.: Using eye movement activity as a correlate of cognitive workload. *Int. J. Ind. Ergon.* **36**(7), 623–636 (2006)
14. Van Orden, K.F., Jung, T.P., Makeig, S.: Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biol. Psychol.* **52**(3), 221 (2000)
15. Long, G.M.: Iconic memory: a review and critique of the study of short-term visual storage. *Psychol. Bull.* **88**(3), 785 (1980)
16. Sewell, D.K., Lilburn, S.D., Smith, P.L.: An information capacity limitation of visual short-term memory. *J. Exp. Psychol. Hum. Percept. Perform.* **40**(6), 2214 (2014)
17. Raymond, J.E., Shapiro, K.L., Arnell, K.M.: Temporary suppression of visual processing in an RSVP task: an attentional blink? *J. Exp. Psychol. Hum. Percept. Perform.* **18**(3), 849 (1992)

Cognition in Aviation and Space

Study on the Astronaut Error Criteria of a Manually Controlled Rendezvous and Docking Operation

Jiayi Cai¹, Weifen Huang^{1,2(✉)}, Jie Li¹, Wang Liu¹, Haipeng Jing¹, Dong Chen¹, Yanlei Wang¹, and Xiang Zhang¹

¹ China Astronaut Research and Training Center, Beijing 100094, China
hwf_2006@sina.com

² National Key Laboratory of Human Factors Engineering,
China Astronaut Research and Training Center, Beijing 100094, China

Abstract. *Objective:* In this paper, some manual control rendezvous and docking operations were researched, astronaut manual rendezvous and docking cognitive decision-making process and its operational characteristics were analyzed, and then the manual rendezvous and docking operations mistakes criteria was determined. *Method and Result:* By capturing operation errors during the operations, it can be easy to find weak points in astronaut's training to provide future reference. *Conclusion:* It can accumulate the basis data for further process optimization rendezvous and docking procedure evaluation methods.

Keywords: Manual rendezvous and docking · Operator errors · Criterion · Training

1 Introduction

Manually controlled rendezvous and docking refers to the operation in which an astronaut observes the relative positions and attitudes of two space vehicles through shipborne equipment such as a television camera on the tracing space vehicle, and then operates the manual controller to perform rendezvous and docking [1]. Manually controlled and automatic rendezvous and docking technology serve as backup for each other [2]. In the man-machine interface system that comprises the astronaut and the spacecraft, the astronaut handles control operations while performing the manually controlled rendezvous and docking operation. In these operations, certain errors are unavoidable. These operational errors are anthropogenic as they result from human error [3].

Human factors analysis and classification system is a qualitative method employed in research on human errors. Specific and targeted human error corrections and preventive measures can be determined through retrospective classification analysis. The classification of human errors may be applied in many fields including spaceflight, nuclear power, transport, and other technically complicated areas [4–7]. Human error classification research has already been performed abroad. Several human error classification systems have been proposed, for example, Norman [8] proposed promulgating

the seven-stage action theory of human behavior in man–machine interactions, Reason [9, 10] proposed human apparent error and human hidden intention error classification models, and Rasmussen [11, 12] proposed the error classification method of knowledge-based, rule-governed, skill-oriented, cognitive behavior models, among other classical classification theories.

Currently, operational errors in manually controlled rendezvous and docking performed by astronauts are not clearly defined. In this paper, the evaluation criteria mainly pertained to handling operational errors based on the abovementioned human error classification models and the practical training experience of the astronaut. By analyzing manual control rendezvous and docking operations, operational errors were classified as relative spatial relationship perception errors, control decision errors, and handle operation execution errors based on cognitive psychology. As these operational errors will increase the number of repetitions of an operation in the entire manually controlled rendezvous and docking process, it can result in increased fuel consumption. Meanwhile, the in-flight stability of the spacecraft will also be affected and the failure rate of the docking mission will increase. It is of vital importance to determine a method to recognize operational errors in the training process more accurately and increase the operational accuracy rate of astronauts during training for manually controlled rendezvous and docking missions. By laying down criteria to capture error classifications effectively, this paper provides support for cognition, decision process, and the design of corresponding software for capturing operational errors made by astronauts.

2 Operational Definition

Astronauts accomplish rendezvous and docking of two spacecraft through several operations using the control handle. Each operation includes the processes of perception, which pertain to the perceived relative spatial orientation of the two spacecraft, decisions relating to handle control, and implementation of actual operational behavior. The determination of relative spatial orientation provides the basis for control handle decisions necessary for further control and implementation of the corresponding handle operation. The result of each operation will serve as feedback to provide new status information for determining the subsequent spatial positioning of the spacecraft.

The procedures for image perception, control decision, and operational implementation are conducted continuously and consecutively in the manually controlled rendezvous and docking mission until the docking process between the two spacecraft is complete. Thus, each control handle operation used for the control of the rendezvous and docking process in the astronaut's manual was studied to decide the corresponding operational error criterion.

The astronaut performs manually controlled rendezvous and docking by controlling the handle for translational motion and the attitude handle to complete the operation. Each time the astronaut controls the handle, a corresponding output voltage is generated. The thrust size and direction of the spacecraft propulsion system are determined based on the direction and size of the output voltage of the handle. Thus, the operational motion of the astronaut can be recognized by the change in output voltage of the

handle during the docking process. Therefore, each operation was defined as the process from the start of operation of the control handle (for translation or attitude) to the end stage where production of voltage signal from the handle is reset to zero, which is the voltage signal at the initial position.

3 Handle Control Characteristic and Operation Analysis

The two handles in rendezvous and docking equipment have different control characteristics. Thus, the operation strategies are also slightly different. Therefore, it is necessary to perform classification analysis on the observed errors.

3.1 Translational Handle Control Process and Error Analysis

The translational handle has two handle heads, one of which can be propelled forward and backward to control the translational motion of the spacecraft in the forward and backward directions. The big-head handle can have translational motion along a single axis or both axes to separately or simultaneously control the translational motion of the spacecraft in the downward, left, or right directions.

During manually controlled rendezvous and docking, the translational handle can change the thrust direction. In addition, the size of the spacecraft influences its speed. When the translational handle returns to zero, the speed of the spacecraft is unchanged. When the acceleration returns to zero, the spacecraft maintains an approximately constant forward speed. This requires the astronaut to prejudge an appropriate opportunity to slow down or speed up to maintain the spacecraft in an appropriate position [10].

The translational position reflects the relative position and distance between two spacecraft. The astronaut estimates the relative position information between two spacecraft using the image information. Later, the astronaut controls the handle to make the engines of the spacecraft to produce higher speeds. At a remote distance, positional deviation can be estimated using information such as the lateral visible area of the target spacecraft. At a close range, the astronaut can estimate the positional deviation by using the distance between the scribed line on the upper chassis and the center of the cross target. In actual operation, the astronaut may incorrectly perceive the current image information, incorrectly estimate the spatial position relationship, fail to consider the relationship between current position and speed during decision-making, or commit control operation errors, all of which can cause the spacecraft to travel beyond the preset position with increasing deviation in translational position. Thus, errors relating to the translational handle operation were regarded as rendezvous and docking operation errors.

3.2 Attitude Handle Control Characteristics and Process Analysis

The attitude handle has three axes of rotation. It can rotate around a single axis, around two axes at the same time, or around all three axes to control the roll, pitch, and off-course attitude.

The attitude relationship reflects the relative attitude angle between two spacecraft, including those for off-course, pitch, and roll. The attitude of the maneuvering spacecraft should be consistent with the attitude of the target spacecraft to complete rendezvous and docking successfully. Therefore, the astronaut must adjust the flight attitude of the spacecraft during the rendezvous and docking process. With regard to the judgment of attitude, the astronaut needs to reconstruct the spatial relationship between the two spacecraft based on the spatial image, which requires imagining the three-dimensional relative attitude relationship using the two-dimensional frame information. Thus, the attitude relationship of the spacecraft must rely on accurate imagination and judgment regarding three-dimensional space of the two spacecraft to allow correct adjustments. During the actual operation, it is easy to err in perceiving current image information, misjudge the attitude direction relationship, operate the wrong handle when making decisions, or perform an incorrect operation while controlling the handle; these errors result in a large deviation in spacecraft attitude direction and high fuel consumption, which prevent successful docking. Therefore, attitude handle operation errors were regarded as rendezvous and docking operation errors in this paper.

3.3 Field Switching Control Characteristics and Process Analysis

The astronaut observes the target spacecraft mainly through the image information captured by wide field and narrow field cameras during the process of rendezvous and docking. The field angle of the wide field camera is large, which makes it convenient for the astronaut to observe the target spacecraft from a longer range and acquire comprehensive information of the docking channel. The field angle of the narrow field camera is small, which makes it convenient for the astronaut to observe the aim-point and other information regarding the target spacecraft clearly to accomplish rendezvous and docking accurately. In an actual docking operation, when the maneuvering spacecraft is at a remote distance, the astronaut can switch to the narrow field camera to acquire detailed image information at the time when the target spacecraft appears at the center of the display screen that displays the frame captured by the wide field camera. In engineering design, the position of the wide field camera can deviate from the target position of the spacecraft. Over-reliance on the wide field camera can present a risk of failure during docking.

4 Design of Operational Error Criteria

Based on the above-mentioned analysis, manual control rendezvous and docking operation errors were divided into three categories: translational handle operation errors, attitude handle operation errors, and field switching operation errors.

4.1 Operational Error Criteria for Translational Handle

With the objective of determining the translational handle control characteristics, translational handle operation errors were considered to exist under the following four conditions:

Error in Control Direction. The operator makes a control action in the opposite direction, which results in increased deviation between the spacecraft and the target location. For example, when the spacecraft is moving leftward at a certain speed and the operator moves the handle towards the left to make the spacecraft accelerate further leftward, it can result in the spacecraft deviating from the target spacecraft at a higher speed.

Insufficient Degree of Operation. The current speed of the spacecraft and relative deviation between it and the target spacecraft are misjudged resulting in a low degree of control. For example, when the spacecraft is moving toward the target point and the astronaut operates the handle to slow down the spacecraft, an insufficient degree of control can cause the spacecraft to overshoot its intended target position.

This error can occur in two circumstances:

1. The perception of speed is wrong. For example, when the spacecraft is moving close to the target spacecraft and the braking mechanism is insufficiently applied, the spacecraft will not be able to slow down and stop at the target point.
2. The perception of position is wrong. For example, when the spacecraft is moving close to the target point, an insufficient deceleration due to miscalculation of position can cause the spacecraft to overshoot the target point.

Excessive Degree of Operation. The degree of control may be in excess if the operator misjudges the current speed of the spacecraft and the relative deviation between the targets. For example, when the spacecraft is moving towards the target point, the astronaut operates the handle to accelerate the spacecraft to get it closer to the target point. An excessive degree of operation causes the spacecraft to overshoot the target point.

The operator may also incorrectly perceive the position. For example, an excessive amount of control operation causes the spacecraft to overshoot the target point.

An incorrect perception of speed can also lead to an excessive amount of control operation. For example, when the spacecraft is too close to the target spacecraft and excessive acceleration is applied, the spacecraft will not be able to slow down sufficiently to stop at the target point.

Untimely or Omitted Operation. When speed of the spacecraft or relative deviation between the targets are misjudged, the control handle is not operated in time, the spacecraft comes close to the target spacecraft at a high speed, or direct deviation between the two spacecraft is small, the maneuvering spacecraft will overshoot the target spacecraft if there is a lack of timely manipulation of speed.

Operational Error Classification for Translational Handle

Error 1: direction operation error

Error 2: perception of speed was wrong, which resulted in excessive operation

Error 3: perception of position was wrong, which resulted in excessive operation

Error 4: perception of speed was wrong, which resulted in insufficient operation

Error 5: perception of position was wrong, which resulted in insufficient operation

Error 6: untimely or omitted operation

4.2 Operational Error Criteria for Attitude Handle

With the objective of determining the attitude handle control characteristics, attitude handle operation errors were considered to occur in the following two circumstances:

Control Direction Error. If the current relative attitude of the spacecraft is misjudged, it can result in wrong adjustments of direction when controlling the attitude handle. For example, when the spacecraft is in left drift, the astronaut may operate the handle in a wrong direction resulting in an increase in the angle of the left drift.

Excessive Control Operation. Because the relative attitude between two spacecraft is usually very small in actual manually controlled rendezvous and docking, the spacecraft will have an angular speed of forward and backward drift. The attitude of spacecraft can be changed by controlling the handle. Thus, accurate attitude operation is critical for operational quality. In this study, we focused on errors caused by the operator because of incorrect perception, incorrect decision, or incorrect operation. Therefore, we mainly aimed to capture the errors relating to the direction of attitude control. The control of the attitude angle was judged as an operational error only if the adjustment made to the attitude angle was no larger than the deviation of the initial adjustment.

Operational Error Classification for Attitude Handle

Error 1: error in direction of operation

Error 2: error in degree of operation

4.3 Field Switching Operational Error

According to the requirements mentioned in the manual control rendezvous and docking training, docking must be completed during the time when the narrow field camera is being used. For the convenience of implementing follow-up operations, it is required that the astronaut should stabilize the target aircraft to the center of the screen when there is a distance of 30 m between the two spacecraft. Therefore, 30 m was selected as the specified distance by which it is necessary to have switched to the narrow field camera. Not switching to the narrow field camera at $X \leq 30$ m was defined as an operational error.

Field Switching Operational Error Classification

Error: narrow field switching was not conducted when $X \leq 30$ m

5 Implementation of Operational Error Capture Software

Operational error capture software can identify incorrect operations made by astronauts during manually controlled rendezvous and docking. In addition, the perception and decision-making process of astronauts were analyzed using a complementary questionnaire. The system included two major modules: the data access module and the error treatment module.

The data access module is responsible for reading the initial data from the rendezvous and docking simulator and generating output data after the errors are corrected. According to the definition of the manually controlled rendezvous and docking operation, the operation begins from the absolute value of the voltage signal received by the manually controlled rendezvous and docking simulator, which is 0.5 V, and ends when the voltage signal returns to zero again. This is marked as a manually controlled rendezvous and docking operation unit of the astronaut. During the entire task execution process, the data access module extracted the operation units for error analysis from the simulator continuously. The output process consisted of error judgment, extraction of error operation by treatment module, error classification information, and generation of error questionnaire module. The error analysis for perception and decision making errors was output to an Excel file. For convenience of further data analysis, the output content includes operator ID, starting time of the task, total time, generated error operation number, relative flying speed of the spacecraft when an incorrect operation occurs, relative translational deviation, off-course attitude, direction of incorrect operation, and value of generated voltage.

The error treatment module is the essence of the software, which conducts error identification and classification according to the error judgment procedure performed on manually controlled rendezvous and docking operational data of astronauts. According to the operation units extracted by the data access module, the error judgment was conducted using a combination of error criteria and status information on the operation units obtained from the simulator. If it was determined that the operation units were in error, the error types were classified. In addition, the resulting information regarding error type, error quantity, and error rate (error quantity/total operation units) were transmitted to the data access module.

6 Software Testing

The operational error capture software had to be tested to verify its accuracy. In this study, primary instructors from the Chinese Astronaut Center were recruited to test various types of errors by designing specific test cases.

In the above-mentioned error criteria, the categorization of manually controlled rendezvous and docking operation errors and the defined error types included the procedures for translational handle, attitude handle, and field switching. There were nine types of errors in total. In the actual operation, the circumstances under which these nine types of errors occurred were recorded. These included the 9 error types corresponding to 37 error circumstances and 20 proper operational circumstances. When designing test cases, all 57 correct and error circumstances were taken into consideration. In this paper, 120 operational test cases were designed, which were aimed at the above-mentioned 57 operational circumstances. After testing all cases, the software met the criteria specified in the design requirement and judgment accuracy of the error software was 100%.

After the testing was concluded, it was replayed through a video. The accuracy and coverage of error identification were estimated by spaceflight experts. It was discussed and reviewed by astronauts and the engineering department. It was concluded that the

design of the error capture software was satisfactory, which made the software capable of detecting operational errors in the manually controlled rendezvous and docking process for the benefit of the astronaut. The criteria used were reasonable but not excessively strict, which can be utilized for the detection of manually controlled rendezvous and docking operation errors.

7 Summary

In this study, the currently available classification models of human errors were reviewed. The operational characteristics of manually controlled rendezvous and docking process and the operational decision-making process of the astronauts were analyzed using cognitive psychology and based on manually controlled rendezvous and docking training given to astronauts. The same training guidelines were used to develop the criteria for rendezvous and docking operational errors, which were used for detecting errors in the perception stage, decision-making stage, and operational performance stage of the astronaut. Depending on the operational axis, operational errors were classified into translational handle control errors, attitude handle control errors, and field switching operational control errors. According to the cognitive process, the operational errors were classified into spatial perception errors, control decision-making errors, and handle operation errors. During their compilation, the error criteria were reviewed and tested by training instructors and astronauts. The accuracy of the error criteria was ensured by creating a detailed design and by modifying each error condition. This laid the foundation for further improvement of the error capture software. Additionally, this has provided a reference for evaluation of rendezvous and docking and improved the astronaut training process.

References

1. Zhiqiang, T., Chunhui, W., Dongxu, H.: Manual control rendezvous and docking ergonomics requirement and evaluation method study. National Defense Science and Technology Report of China (GF-A0129991M), China Astronaut Scientific Research Training Center (2010)
2. Fehse, W., (Dongxu, L., Zhi, L. (transl.)): Automated Rendezvous and Docking of Spacecraft. Press of National Defense Science and Technology University (2008)
3. Swain, A.D., Guttman, H.E.: Handbook of human reliability analysis with emphasis on nuclear power plant applications, Final report, NUREG. CR-1278, Sandia National Laboratory (1983)
4. Amalberti, R., Wioland, L.: Human error in aviation. In: Aviation Safety, pp. 91–108 (1997)
5. Bell, B.J., Swain, A.D.: A Procedure for Conducting a Human Reliability Analysis for Nuclear Power Plants. US Nuclear Regulatory Commission, Washington (1983)
6. Van der Schaaf, T.W.: Misreporting in the chemical process industry. Doctoral thesis. Eindhoven University of Technology, Eindhoven (1992)

7. Stanton, N.A., Salmon, P.M.: Human error taxonomies applied to driving: a generic driver error taxonomy and its implications for intelligent transport systems. *Saf. Sci.* **47**(2), 227–237 (2009)
8. Norman, D.A., Draper, S.W.: *User Centered System Design*. Lawrence Erlbaum Associates, Hillsdale (1986)
9. Reason, J.: *Human Error*. Cambridge University Press, New York (1990)
10. Meister, D.: The Nature of Human Error, pp. 783–786. Naval Ocean Systems Center, San Diego (1989)
11. Rasmussen, J.: Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Trans. Syst. Man Cybern.* **3**, 257–266 (1983)
12. Rasmussen, J.: Human errors. A taxonomy for describing human malfunction in industrial installations. *J. Occup. Accid.* **4**(2), 311–333 (1982)
13. Wang, C., Ting, J.: Ergonomics study on display-control system of manual control rendezvous and docking task. *Manned Spaceflight* **17**(2), 50–53 (2011)
14. Jin, Y., Guohua, J., Jiangang, C.: Astronaut manual control rendezvous and docking method based on target image. *J. Astronaut.* **31**(5), 1398–1404 (2010)
15. Yanlei, W., Xiang, Z.: Control strategy research report of astronaut manual control rendezvous and docking V2.0. Technical Documents of Astronaut Scientific Research Training Center [internal] HYK-X000GZ41 (2012)

Multi-modal Interaction Between Pilots and Avionic Systems On-Board Large Commercial Aircraft

Jason Gauci¹(✉), Matthew Xuereb¹, Alan Muscat²,
and David Zammit-Mangion¹

¹ Institute of Aerospace Technologies, University of Malta, Msida, Malta

{jason.gauci, matthew.xuereb,
david.zammit-mangion}@um.edu.mt

² QuAero Ltd., Mosta, Malta

alan.p.muscat@quaero.aero

Abstract. A lot of work has been carried out over the last decade to apply touchscreen technology to the flight deck of large commercial passenger aircraft. In fact, several industry solutions are now available and aircraft equipped with touchscreen solutions will be flying in the very near future. In contrast, Direct Voice Input (DVI) technology is still several years away from entering commercial service on large transport aircraft; nevertheless, it has a lot of potential and can even overcome some of the limitations associated with touchscreen technology. This paper presents a prototype application based on DVI which enables pilots to interact with the autopilot by means of voice commands. This application is composed of a speaker dependent speech recognition module and a command recognition module. The results of an evaluation of the DVI application are presented and discussed in detail and areas of improvement are outlined. The DVI application is part of a bigger solution which is intended to combine the benefits of touchscreen technology and DVI into a single multimodal interface.

Keywords: HMI · Touchscreen · DVI · Cockpit · Multimodal

1 Introduction

To this day, pilots of large commercial aircraft – including those produced by Airbus and Boeing – still interact with on-board avionic systems using conventional interfaces such as the Flight Control Unit (FCU)¹ for autopilot control, the Control and Display Unit (CDU) for flight management, and the Radio Management Panel (RMP) for communication and navigation frequency selection. Information about the current state of the aircraft and its various systems is shown on a number of instrument displays such as the Primary Flight Display (PFD) and Navigation Display (ND).

A lot of work has been carried out over the last decade to introduce new modes of interaction to the flight deck, such as touchscreen technology. There are several advantages associated with this technology, such as the ability to manage avionic

¹ Mode Control Panel (MCP) in the case of Boeing aircraft.

systems in a more intuitive manner, and to control systems and view their status from the same display. Touchscreen solutions are already beginning to appear on business jets [1] and will soon be introduced to commercial passenger aircraft as well [2].

A number of research projects have explored the application of touchscreen technology to the flight deck [3–6]. In [6] the University of Malta developed a concept for tactical and strategic flight control which is based on the use of a touchscreen device which can be placed on the table in front of each pilot. With this device, each pilot can interact with multiple avionic systems through a single interface without having to reach out to different controls located around the flight deck. This can be particularly advantageous when flying through turbulence. An image of the touchscreen interface is shown in Fig. 1.



Fig. 1. The touchscreen interface concept

Another mode of interaction which can be introduced to the flight deck is voice control (also known as Direct Voice Input (DVI)). Voice control has many potential benefits; for instance, pilots can issue commands ‘hands-free’, thus allowing them to use their hands for other tasks. Also, pilots do not need to look down at a particular screen and can therefore spend more time looking outside for other aircraft and potential obstacles. On the other hand, voice control presents a number of challenges, including the ability to cope with voice differences (such as accents and intonations) and noise in the cockpit. DVI has been used on military fighter jets (such as the Eurofighter Typhoon) for several years [7] but is yet to be introduced on commercial aircraft. In [8] the authors explore the use of multiple modes of interaction – including touch and voice control – for the purpose of flight management.

This paper focuses on an ongoing research project which builds on the results of [6] by extending pilot interaction with cockpit avionics through the use of DVI. This will enable pilots to use voice commands and/or touchscreen gestures to control the aircraft. One of the advantages of DVI that is exploited in this project is the possibility to recognize commands that are issued by the pilot when acknowledging (repeating)

instructions or clearances issued by Air Traffic Control (ATC).² For instance, if ATC requests the crew to fly on a certain heading, the pilot will repeat the heading instruction back to ATC and the DVI application will recognize the pilot's heading command. Then, all the pilot has to do is execute the command. Hence, this has the potential to reduce crew workload.

The rest of this paper is organized as follows. Section 2 highlights the main requirements associated with DVI. Section 3 presents the voice commands which were defined for tactical flight control. Section 4 describes the main components of the DVI application. Sections 5 and 6 discuss the preliminary evaluation of the DVI application and Sect. 7 presents the main conclusions of this work.

2 DVI Requirements

In order to ensure that voice control can be applied to the cockpit environment in a way that is acceptable to pilots, while still being compatible with current flight deck procedures and operations, the DVI application needs to meet a number of challenging requirements, including:

Standard phraseology – The wording and structure of the voice commands should be based on standard phraseology that is used between pilots in the cockpit, as well as between the crew and ATC. This reduces the amount of training that would be required for pilots to use the DVI application, and makes it easier to recall specific instructions.

Flexibility – The application should be able to cope with common variations of the same word or command. For example, the number 250 in the command 'SET SPEED 250' can be expressed in different ways such as 'TWO HUNDRED FIFTY', 'TWO FIFTY' and 'TWO FIVE ZERO'. Similarly, certain words may be omitted from a command or the order of words may differ, without affecting the meaning of the command.

Robustness and recognition accuracy – The application should have a high speech recognition rate (of at least 98%) and should be able to cope with the noise levels of a typical flight deck environment. Ideally, the application should be speaker independent such that it can recognize different voices and cope with variations in accent, pitch, intonation and other speech characteristics. This is important since flight crew members change between flights. Alternatively, if a speaker dependent solution is used, users need to log onto the system in order to identify themselves and ensure that the correct voice profile is used by the application.

Response time – The DVI application should recognize commands in a timely manner (ideally within 200 ms) and provide adequate visual and/or aural feedback to the user. This is especially important in critical situations and when the pilot needs to react quickly to ATC instructions.

Safety – Depending on the nature of the command and the criticality of the situation at hand, a voice command may need to be confirmed by the pilot prior to execution.

² This is known as ATC read back.

This confirmation introduces a delay but may be necessary to ensure that the correct command will be executed. The research project described in this work is investigating whether command confirmation should be mandatory in all cases or not. In order to enhance safety, the DVI application should also validate the voice commands and check that they are consistent with the current state of the aircraft. The user should be informed if an invalid command is detected.

3 Voice Commands for Tactical Control

Voice commands were defined for various aspects of tactical flight control (i.e. autopilot control via the FCU), including: switching the autopilot ON or OFF, engaging different autopilot modes, and selecting target values for speed, heading, altitude and vertical speed. A non-exhaustive list of voice commands for autopilot control is presented in Table 1 together with some examples which demonstrate the command syntax used.

Table 1. Voice commands for autopilot control

Function	Description	Command
Autopilot ON/OFF	Switch the first autopilot ON	AUTOPILOT ONE ON
	Switch the second autopilot OFF	AUTOPILOT TWO OFF
Speed mode	Engage Selected speed mode	SELECT SPEED
	Engage Managed speed mode	MANAGE SPEED
Target speed	Set speed in knots	SET SPEED 2-50 KNOTS
		SPEED 2-5-0
		SET 2-50 INDICATED
		INCREASE 2-50 KNOTS
	Set speed in Mach	SET SPEED DECIMAL-7-8
Maintain present speed	MAINTAIN PRESENT SPEED	
Heading mode	Engage Selected heading mode	SELECT HEADING
	Engage Managed heading mode	MANAGE HEADING
Target heading	Specify absolute heading	SET HEADING 1-3-5 DEGREES
		SET HEADING 1-3-5
	Set heading to a cardinal compass value	HEADING NORTH
	Specify absolute heading and direction of turn	TURN LEFT HEADING 1-3-5
		LEFT HEADING 1-3-5
Specify relative heading change	LEFT 10 DEGREES RIGHT 90 DEGREES	
Vertical mode	Engage Managed climb mode	MANAGE CLIMB
	Engage Open climb mode	OPEN CLIMB
	Engage Managed descent mode	MANAGE DESCENT
	Engage Open descent mode	OPEN DESCENT

(continued)

Table 1. (continued)

Function	Description	Command
Target altitude	Set altitude in feet	SET ALTITUDE 5000 FEET
		SET ALTITUDE 5000
		STOP CLIMB 5000 FEET
	Set altitude in Flight Level (FL)	SET FLIGHT LEVEL 0-9-0
		DESCEND FLIGHT LEVEL 9-0
Level off the aircraft	LEVEL OFF	
Stop the descent	STOP DESCENT NOW	
Target vertical speed	Descend at a particular vertical speed	DESCEND 1500 FEET PER MINUTE
		DESCEND AT 1500 FEET PER MINUTE OR MORE
	Continue climb/descent at current vertical speed	SELECT VERTICAL SPEED
Upset recovery	Level off aircraft and maintain current heading and speed	STOP
Confirmation	Confirm and execute command	EXECUTE

From Table 1 it can be observed that there are various instances where the exact same function can be performed using different voice commands. For instance, when setting an aircraft speed of 250 knots, one of (at least) four different commands can be used. This is representative of the variations that can be found in practice. It can also be observed that, apart from setting the value of an autopilot parameter, the user can also engage autopilot flight modes. For instance, in the case of heading and speed, the user can switch between Selected mode (in which case the autopilot follows a user-selected heading or speed) and Managed mode (in which case the autopilot follows the FMS plan). In the case of altitude, the user can engage an Open or Managed climb/descent mode. For a Managed climb or descent, the autopilot follows the FMS plan. For an Open climb, the auto-throttle is set to full climb thrust whereas, for an Open descent, the auto-throttle is set to idle thrust. Furthermore, any constraints on the FMS plan are disregarded during an Open climb or descent.

The 'EXECUTE' command is issued by the pilot after each of the voice commands shown in Table 1. This enables the pilot to confirm that the voice command has been correctly identified by the DVI application.

4 DVI Application

The DVI application is composed of two main modules, the first of which is a speech recognition module. For this project, the commercially available Dragon Naturally-Speaking Premium 13.0 voice recognition engine by Nuance[®] was used. This is a speaker dependent speech recognition system which uses advanced machine learning

algorithms (such as Deep Neural Networks (DNN)) to recognize speech and is able to adapt and learn in order to improve its recognition accuracy.

The speech recognition software comes with a default dictionary that can be updated by the user. It also has a training sub-module that can be used to train the software to recognize individual words. In fact, during the development of the DVI application, the speech recognition software was trained to recognize each of the words used in the voice commands. Furthermore, the dictionary was modified by removing words which sounded similar to those used in the voice commands. This was done in an attempt to improve the recognition accuracy of the software.

The second main module of the DVI application is a custom-built software package which processes the output of the speech recognition module in order to identify specific voice commands. This is essentially done by checking whether the words, format and syntax of phrases spoken by the user match one of the predefined voice commands for autopilot control. If a match is found, the software checks the validity of the command, such as by checking that any numerical values are within a certain range, depending on the current aircraft configuration. For instance, the upper and lower limits of target speed vary dynamically with flap setting. The software also checks for inconsistencies in the commands. For example, if the user issues the command 'DESCEND FLIGHT LEVEL 100' but the aircraft is at flight level 50, an inconsistency is detected.

The DVI module is activated (and therefore listens to voice commands) only while the user presses the Push-to-talk (PTT) button. The speech that is detected by the application is displayed to the user. If the voice command is invalid, the user is informed by means of the error message 'INVALID COMMAND'. If the command is valid but the parameter being set (e.g. speed) exceeds certain limits, the error message 'INVALID RANGE' is shown. On the other hand, if the command is recognized by the system and is valid, it is read back to the user via the headset. The user can then confirm the command by saying 'EXECUTE'. This will trigger the DVI application to transmit the command to the flight simulation platform. In this research, X-Plane is used as the flight simulation environment and the aircraft model is the Airbus 320 New Engine Option (A320neo).

The voice commands are defined in a script file, together with the associated actions and any preconditions and/or constraints. The user can easily modify this script file without having to recompile the source code. The following is a small section of the script file which defines commands for setting an absolute heading:

```
TargetHeading, #x#
#x#
Range: 0..360
Type: integer
{
    SET HEADING #x# DEGREES
    SET HEADING #x#
    HEADING #x#
    HEADING #x# DEGREES
}
```

In this sample of the script file, four variations of the same heading command are defined between curly brackets. The target heading is represented by '#x#' and is constrained to integers between 0 and 360. If the command is valid and is confirmed by the user, the target heading is loaded into the variable Target Heading which is transmitted to the A320neo model in X-Plane. The target heading will then appear on the simulator's FCU and PFD and the aircraft will start turning towards that heading.

5 Evaluation

In order to get user feedback on the DVI application before extending it to other functional areas (including the FMS and communication system), a preliminary evaluation was carried out with a number of commercial airline pilots, with one pilot per evaluation session. Each evaluation consisted of a briefing, an acclimatization session, some test scenarios, and a debriefing.

During the briefing, the pilots were first given an overview of the scope and objectives of the project and the evaluations. Then they were asked to complete a short questionnaire about their flying experience. After the briefing, the pilots were shown how to use the DVI application; then, they were provided with a headset and told to train the speech recognition software by reading out loud a list of words which formed part of the voice commands. Following that, they were allocated some time to practice giving the application voice commands.

The main part of the evaluation session consisted of a number of test scenarios where the evaluation pilot was given instructions by one of the researchers who acted as a pseudo Air Traffic Controller (ATCo). These included instructions to adjust the aircraft speed, heading, altitude and vertical speed. After each instruction, the evaluation pilot had to read the instruction back to the ATCo (according to standard procedure) while pressing the PTT button, thereby activating the DVI application. If the application recognized the instruction correctly and considered it to be valid, the pilot had to confirm it and then the command was executed.

Following the test scenarios described above, the evaluation pilot was asked to close his eyes while the aircraft was initialized in an unusual attitude. The pilot was then asked to open his eyes and recover the aircraft using voice commands. This test was repeated twice. When the tests were complete, a debriefing was carried out and the evaluation pilot was asked to complete a questionnaire related to various aspects of the DVI application.

During the evaluations, qualitative data was gathered by means of questionnaires, video recordings, semi-structured interviews and direct observation.

6 Results and Discussion

Three professional civil air transport pilots flying part 25 certified aircraft (Airbus A319/A320/A321) participated in the evaluations. Their flying experience ranged from six years to over 20 years and their age ranged from 33 to over 40 years. Two of the pilots were male first officers while the third pilot was a female captain. Only one of the



Fig. 2. The pseudo ATCo (left) and evaluation pilot (right) during one of the evaluation sessions

pilots had prior experience of using applications based on voice recognition. A photo showing the pseudo ATCo with one of the pilots is presented in Fig. 2.

In general, the pilots agreed with the idea of having voice control in the cockpit as an additional mode of interaction. Two of the pilots said that this would be particularly useful in abnormal or critical situations, such as in the event of smoke in the cockpit. When asked to rate the applicability of voice control to particular avionic systems and functions, the pilots assigned the ratings given in Table 2. From this table it can be observed that the pilots felt that voice control is most suited to the FMS and the communications system, and least suited to the manipulation of flight controls.

Table 2. Ratings assigned by the evaluation pilots for the application of voice control for different systems or functions (1 = totally disagree, 5 = fully agree)

System/function	Rating			Total rating
	Pilot 1	Pilot 2	Pilot 3	
Autopilot	3	4	2	9
FMS	4	4	4	12
Communications system	4	4	4	12
Checklist completion	4	3	2	9
Control manipulation (e.g. flaps, gear lever, speed brakes)	3	2	2	7

In the case of the FMS, one of the reasons for the high score could be the fact that, with current flight deck interface technology, flight plan modifications (such as the addition of a waypoint) require the pilot to navigate through various menus and press multiple buttons on the CDU. In contrast, with voice control, the same operation could potentially be performed with a single voice command. One of the pilots suggested that voice control could also be used to transmit messages via the Aircraft Communications Addressing and Reporting System (ACARS).

The pilots were also asked to rate various aspects of the DVI application Table 3. As can be observed from this table, the pilots assigned a score of 3 or more to the majority of the aspects of the DVI application, with the recognition accuracy and responsiveness of the application scoring the least. The following paragraphs explain some of the reasons behind these scores and discuss potential solutions to improve the performance of the DVI application.

Table 3. Ratings given by the evaluation pilots for various aspects of the DVI application (1 = poor, 3 = acceptable, 5 = excellent)

Aspect of DVI application	Score			Total rating
	Pilot 1	Pilot 2	Pilot 3	
Phraseology and structure of voice commands	3	3	5	11
Accuracy of voice and command recognition	3	3	3	9
Speed of command recognition (i.e. responsiveness of the application)	4	3	1	8
Textual representation of voice commands for pilot confirmation	4	3	3	10

As explained previously, the speech recognition software had to be trained to recognize the voice of the evaluation pilot before the beginning of the test scenarios. This was done by reading out loud each of the words used in the voice commands and checking that they were correctly identified by the system. As expected, in each case there were a number of words which were either recognized intermittently or not at all. In this case, the first step was to train the speech recognition module by repeating each of the words individually for a number of times. If, after the training, the software was still unable to recognize a particular word (e.g. maintain), the dictionary was modified manually by removing any similar-sounding words (e.g. maintenance) which were not being used in any of the voice commands.

The steps described above improved the recognition accuracy of the application during the acclimatization phase; however, during the actual test scenarios, a number of words were still identified incorrectly. This suggests that the training phase of the speech recognition software was not sufficient and that more time was required to enable the software to adapt to (and learn) the user’s voice.

Another solution to improve the performance of the DVI application is to add context to the speech and command recognition process. For instance, certain groups of words are always spoken together in a particular command (such as ‘OPEN CLIMB’ or ‘FLIGHT LEVEL’). In this case, it is possible to train the speech recognition software to recognize a whole phrase (group of words) rather than individual words. Also, in the case of similarly-sounding words, the DVI application can correctly identify an ambiguous word by taking into account any words which are spoken before or after that word. For example, if the speech recognition software outputs ‘10 LEFT HEADING 1-3-5’ (where ‘TURN’ is incorrectly recognized as ‘10’ due to the similarity between the two words), the DVI application can examine the whole command and determine that ‘10’ is out of context and consequently replace it with ‘TURN’.

Another possible solution to improve the accuracy of speech recognition is simply to use a higher quality microphone with noise-cancelling properties. This would reduce the impact of any background noise on the speech recognition process.

The DVI application is currently designed to wait for pilot confirmation (via the 'EXECUTE' command) before executing a voice command. Although this adds a level of safety to the system, all of the pilots agreed that this feature is not desirable or justified in time-critical situations where quick pilot reactions are essential. For instance, when recovering from a stall (or any other upset condition), the 'STOP' command would be sufficient and should be executed immediately. The pilots also suggested that the wording could be changed from 'STOP' to 'RECOVER' or 'LEVEL OFF' since the intention of either of these commands was clear and unambiguous.

The DVI application rated quite well in terms of phraseology; however, the pilots felt that this aspect could be improved by allowing for more flexibility during read back of ATC instructions. One of the pilots also suggested that, when the command is read back to the pilot by the DVI application prior to execution, the phraseology used could be similar to that of the Flight Mode Annunciator (FMA)³. This would make it easier for the pilot to confirm that the correct command will be executed.

An issue that was occasionally observed during each of the evaluation sessions was that pilots either forgot to press the PTT button before issuing a voice command, or released the PTT button too early (i.e. before the text corresponding to the command appeared on the display). One possible solution to this problem is to replace the functionality of the PTT button with a dedicated voice command.

7 Conclusion

This paper presented a prototype application based on DVI technology which enables pilots to interact with the autopilot by means of voice commands. The application was evaluated with the participation of commercial airline pilots and the overall feedback was positive. Several suggestions for improvement were made and a number of potential solutions were identified.

The DVI application is part of a bigger solution that is designed to provide pilots with multiple modes of interaction with avionic systems. The next steps will focus on the application of DVI to other functional areas (including flight management, communication, and checklist execution), the integration of voice control with the touch-screen interface developed in [6], and the evaluation of the complete integrated solution in a representative cockpit environment.

Acknowledgments. The work presented in this paper was carried out as part of TOUCH-FLIGHT 2/ePM, a project financed by the Malta Council for Science & Technology through FUSION: The R&I Technology Development Programme 2016. The authors would like to thank the pilots who participated in the evaluations.

³ The FMA appears on the PFD and is the primary status indicator of auto-flight and auto-thrust modes and engagement status of the auto-pilot, flight directors, and auto-thrust.

References

1. Newest Gulfstream Aircraft Flies with Honeywell Touch. <https://aerospace.honeywell.com/en/news-listing/2015/may/newest-gulfstream-aircraft-flies-with-honeywell-touch>
2. Boeing 777X to feature touchscreen flight displays from Rockwell Collins. http://www.rockwellcollins.com/Data/News/2016-Cal-Yr/CS/FY16CSNR51-777x.aspx?utm_source=rcemail&utm_medium=rcprblast&utm_campaign=FY16CSNR51-777x
3. Final Report Summary – ODICIS. http://cordis.europa.eu/result/rcn/54075_en.html
4. Alapetite, A., Fogh, R., Zammit-Mangion, D., Zammit, C., Agius, I., Fabbri, M., Pregolato, M., Becouarn, L.: Direct tactile manipulation of the flight plan in a modern aircraft cockpit. In: Proceedings of International Conference on Human-Computer Interaction in Aerospace, HCI Aero 2012 (2012)
5. Hamon, A., Palanque, P., André, R., Barboni, E., Cronel, M., Navarre, D.: Multi-touch interactions for control and display in interactive cockpits: issues and a proposal. In: International Conference on Human-Computer Interaction in Aerospace, California, USA, 30 July–1 August 2014 (2014)
6. Gauci, J., Cauchi, N., Theuma, K., Muscat, A., Zammit-Mangion, D.: Design and evaluation of a touch screen concept for pilot interaction with avionic systems. In: 34th Digital Avionics Systems Conference, Prague, Czech Republic, 13–17 September 2015 (2015)
7. Direct Voice Input Technology. <https://www.eurofighter.com/news-and-events/2008/05/direct-voice-input-technology>
8. Dostál, M., Kolčárek, P.: Multimodal navigation display. In: 34th Digital Avionics Systems Conference, Prague, Czech Republic, 13–17 September 2015 (2015)

A Study for Human-Machine Interface Design of Spacecraft Display & Control Device Based on Eye-Tracking Experiments

Qi Guo^(✉), Chengqi Xue, Yun Lin, Yafeng Niu, and Mo Chen

School of Mechanical Engineering,
Southeast University, Nanjing 211189, China
ipd_xcq@seu.edu.cn

Abstract. The display & control device is the hub of human-machine interaction in the whole spacecraft's human-machine environment system. The rationality of its design affects the level of integration between human and machine directly. Therefore, the optimization design studies of manned spacecraft cabin's display & control device plays an important role on the development of spaceflight. The paper researches the design methods of manned spacecraft cabin's display & control device from the perspective of human-machine ergonomics. The specific steps of research are: Firstly, we will refine the design principles of human-machine interface by constructing the user behavior model; Secondly, we will work out the design and layout solutions of spacecraft cabin's display & control device based on the information from user behavior model; Finally, an eye-tracking experiment will be proposed to verify and optimize the solutions.

Keywords: Display & control device · Human-machine interface design · Behavior model · Eye-tracking experiment

1 Introduction

With the development of modern computer technology, the technology of human-machine interaction has become the key theory on the design of display & control device. The design concept of people-oriented is to change the behavior of the machine system into the act of communication between users and machine easily, thus it helps to improve the operation safety, reliability and efficiency. The display & control device is a main part of human-machine environment system, which bridge users and machine information exchange in the environment of human-machine system. British scholar [1] began to study the control panel layout in 1967, multiple computer control panel layout programs were developed (1973 and 1977 respectively), Dr. M.P. Dan [2] put forward computer aided design system (CADS) and expert system (ES), which are applied in the design of ergonomics in the cab of the vehicle. But the method can only be applied to the design of human-machine interface with few components and small layout area, Wei Ning Fang [3] established the constraints between control units and locomotive operation according to the task, function and ergonomics requirements, Wei Liu [4] studied the application of situated cognition in human-machine interaction

design, he pointed out that the study of situational awareness on transportation, intelligent home furnishing, robot intelligence and other related research fields have practical significance. In order to meet the requirements of ergonomics in the design of display & control device, it is necessary to carry out scientific research and experiment, such as the experimental method, the observation method, the subjective questionnaire, the virtual simulation method and the intelligent algorithm model, etc. Due to the limitation of space on the spacecraft, the display & control devices tend to be integrated and systematic. The visual perception and cognitive burden of astronauts are increased because of the excessive display parameters. Therefore, the research on the basic principles of human-machine interaction and physiological characteristics of human beings have become more and more important in the space exploration.

After analyzing the deficiencies of the existing methods of human-machine interface design, the following two steps are proposed to design the layout of display & control device based on physiological and psychological characteristics of astronaut:

1. The human cognition and performance as the two factors of human are the basic research on the information exchange between human and machine. In order to find out the rules of astronaut cognitive and operational behavior, GOMS (goal-operators-methods-selection) [5] is used to analysis the relationship between astronaut factors and interface design, so we can instruct the design of human-machine interface of display and control device.
2. Eye-tracking technology [6] is utilized to analyze and verify the design solution of the manned spacecraft cabin's display & control device. A normalized eye-tracking experiment is arranged to capture eye migration path, fixation point distribution, and the heat map of eye focus [7]. After that, we will find the hot spots of eye focus and the first effect. These parameters are chosen as the qualitative and quantitative evaluation indexes for the human-machine interface design, the experiment results are the important gist to improve and optimize the design solution.

2 Methodology

2.1 GOMS Model for Operating Tasks

GOMS model is applied to the design of human-machine interface for the first time, it plays an important role in the study of the display & control device on spacecraft, it can be used to find out the factors which influence the operation of the display & control device and study the relationship among these factors.

Based on the information processing theory of human problem solving, the GOMS model describes the solving process from four aspects: goal, operator, method and rule. Considering the confidentiality of project, this paper takes the operation process of the communication between the astronaut and the ground command center as an example, and makes the relevant task hypothesis to simulate the interaction between the astronaut and the speech control unit. The GOMS model is used to analyze the operation process of the astronauts and the effect of astronaut's cognitive behavior characteristics on the design of the display & control device, after that, the guidance of interface design will be extracted.

A. Task objectives and their decomposing steps

In this paper, the general objective of the communication task is decomposed into five sub-goals, namely startup, connection, conversation, records and safe operation. Sub-goals can be divided into more specific operational goals. The specific task decomposition is shown in Fig. 1

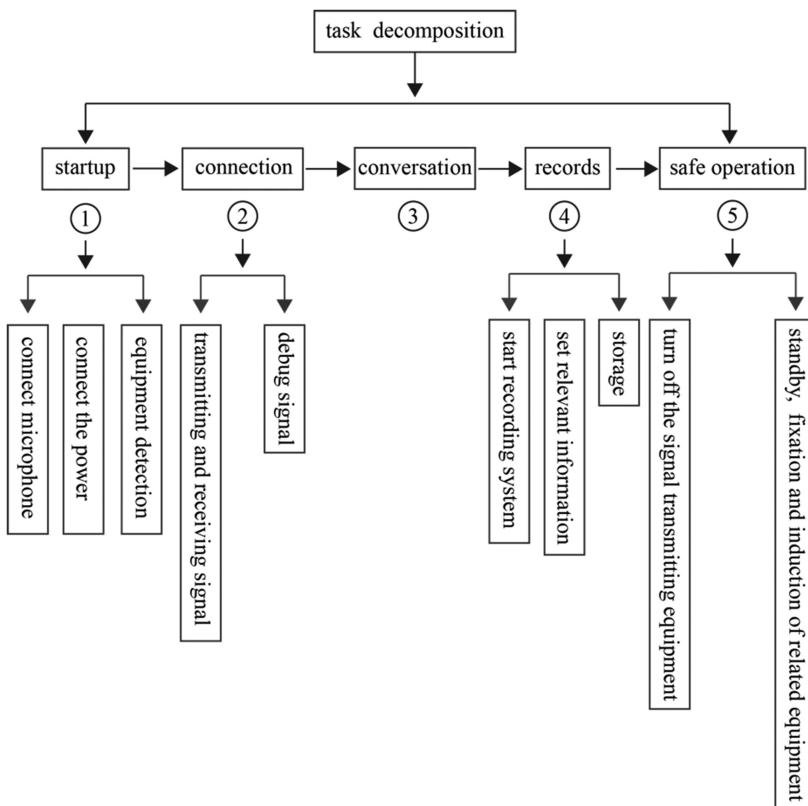


Fig. 1. Specific task decomposition

The following assumptions are made on the model for the astronaut’s information processing:

The perception channel is mainly based on the vision and auditory sense, and the operation mode of the astronaut is the expert mode, namely, there is no error operation. The sub tasks of ①②③⑤ in Fig. 1 are sequential, with the ④ and the subtasks of ③ and ⑤ are simultaneous. Some key switch steps in “record” task: “space suit” - “call” - “volume control” - “recording” - “hang up”. Then import the GOMS model.

The above behavior objectives are introduced into the GOMS model, and the analysis process of some subtasks is shown in Figs. 2, 3 and 4.

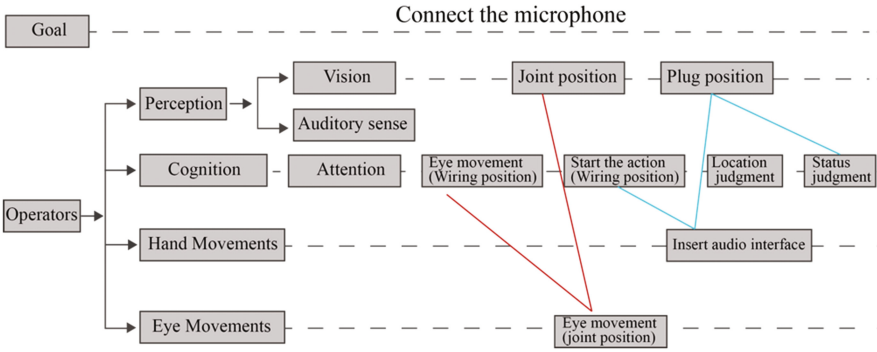


Fig. 2. Operational model of “connect the microphone”

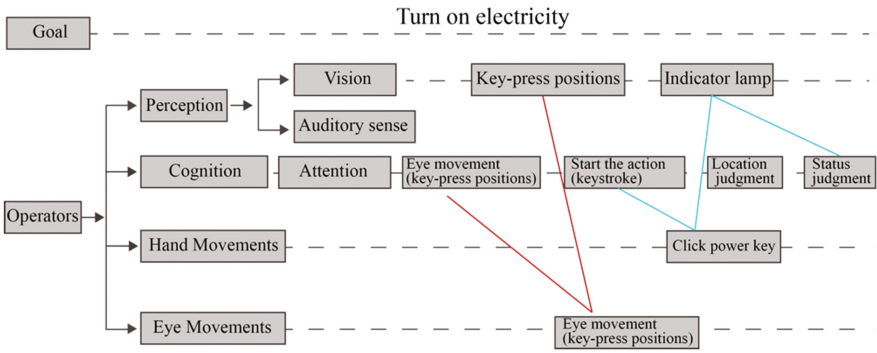


Fig. 3. Operational model of “turn on electricity”

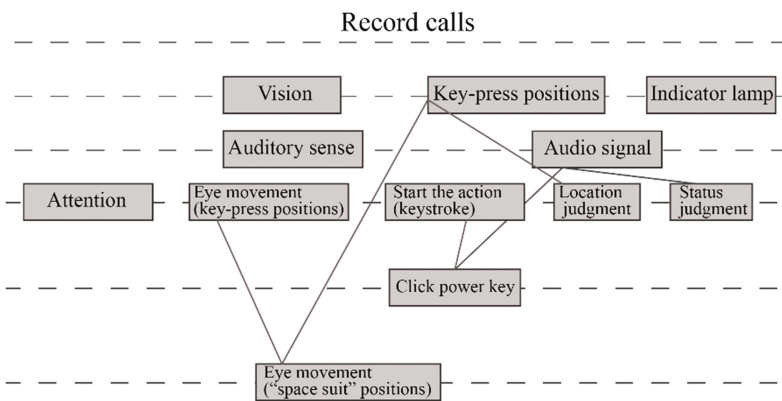


Fig. 4. Operational model of “record calls”

B. Design analysis and guidance

- (1) Analysis of using process: there are many switching actions in the process, therefore, the matching of button position and operation process should be considered.
- (2) Analysis of the environment: voice unit and other display & control device are arranged in the cabin on the operation panel, so the design should be coordinated with the environment of other equipment, including interface tone and overall styling.
- (3) Analysis of the user error: astronauts' error prone to operational errors in the operation process. In order to prevent the situation, the designer will design the feedback result after each step of the operation to carry on the interactive design.

2.2 The Optimal Experiments on the Display & Control Device

2.2.1 Experimental Equipment and Experimental Procedures

Eye movement experiments were carried out with the EyeLink6, integrating and recording the results by the data processing software [8]. Twenty people were elected from the instrumentation engineering, mechanical manufacturing, automation, aerospace manufacturing engineering, aircraft design and other similar professional master of engineering as test subjects. Before the experiment, let the test personnel read experiment notes, and show some different types of display and controller for participants to make a general understanding of the content.

The subjects will observe the two pictures carefully, each picture is equipped with a guide, the first observation object is a blank rectangle with the length and width of 960 mm multiplied by 920 mm, the blank rectangle is divided into nine small areas and four intersections, the guide words are: "please look carefully at the image below the picture, may be round or rectangular, observation time is about 5 s". After the experiment, the fixation points of 20 test subjects will be exported.

2.2.2 The Analysis of Fixation Points

In a blank rectangle, there is no interference of eye gaze movement caused by other visual or cognitive factors. The number of fixation is proportional to the point of interest, the area with the largest amount of attention can be called the golden visual area. The fixation points of the 20 subjects were stacked into a superposition point of view as shown in Fig. 5. It is easy to find that the regions with the highest frequency are the areas marked in Fig. 6, which are the intersection areas, and the measured points were mainly distributed in area A, D, E and F, the area G, H and I below the rectangle is less than the rest.

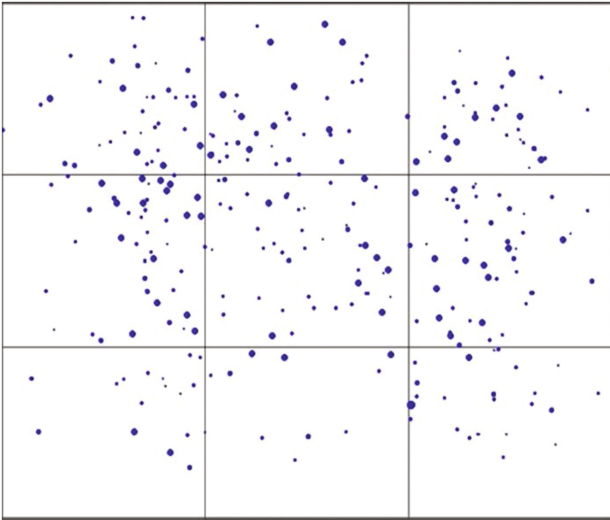


Fig. 5. Superposition diagram

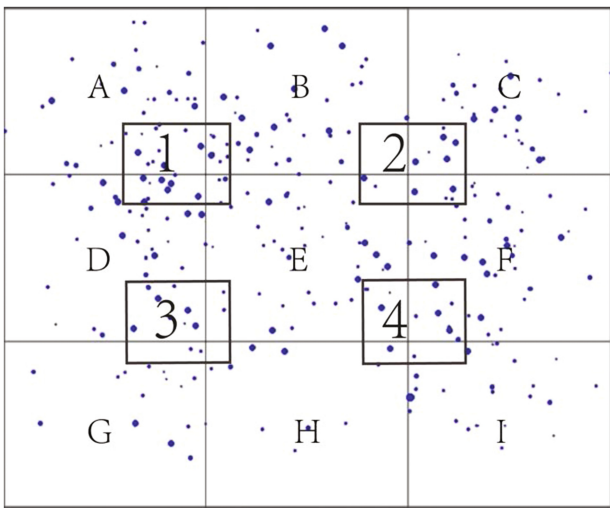


Fig. 6. Fixation distribution

2.3 The Verification Experiment on the Display & Control Device

2.3.1 Experimental Method and Experimental Procedures

The two groups were tested with each group of 20 people. The target scheme and the contrast scheme were selected as the test subject to verify the design of the interface morphology and semantics, the design of the color semantic and the rationality of the layout design. The number of fixation points and the average saccade amplitude were

recorded by the EyeLink6. By calculating the mean and standard deviation of fixation time to evaluate whether the interface design is friendly, by calculating the mean and standard deviation of saccade amplitude, we can evaluate the rationality of the layout design.

In this paper, the design scheme of the experimental test is shown in Fig. 7, in which the A is the target scheme and the contrast schemes are B and C. The experiment was divided into two groups, each group consisted of 20 subjects, the first group of tests were A and B, the second group of tests was A and C.

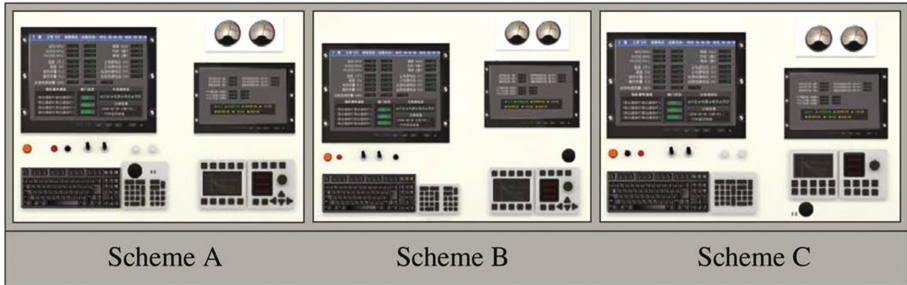


Fig. 7. Experimental scheme

The visual path of the scheme which is shown in Fig. 8.

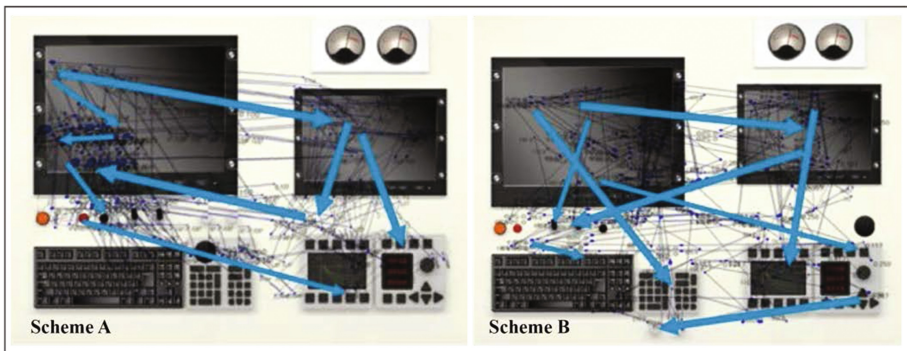


Fig. 8. The diagram of visual path

After analyzing the path characteristics of the eye movement in the task, we can obtain that the visual path of the scheme A is more clear, and the layout of scheme A is in line with the logic habits of human vision. Furthermore, the visual path of the scheme B is more chaotic, it is hard to find the target project except for a large visual jump. Because of the fewer changes of the longitudinal layout, the observational data is not obvious. Therefore, this paper only shows the average saccade amplitude of each tester's interface layout, as shown in Fig. 9.

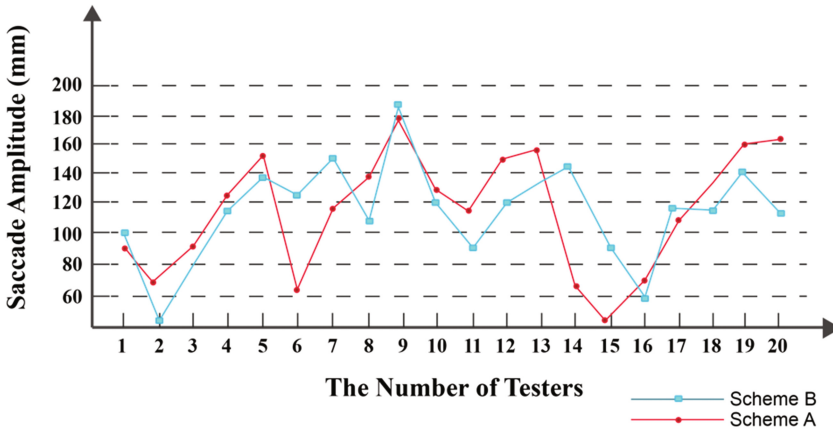


Fig. 9. Average saccade amplitude observed by the first group of testers

The data were analyzed by SPSS, the average saccade amplitude of scheme A is 130.8000 mm, the standard deviation is 34.2416, the average saccade amplitude of scheme B is 114.6794 mm, the standard deviation is 10.1108, it is obvious that the design of scheme A is better than the scheme B.

One-way ANOVA [9] was used to analyze the two sets of data, the original hypothesis is that there is no significant effect on the average saccade amplitude of scheme A and scheme B, let’s suppose that the significant level is 0.05, and the results are shown in Table 1. It indicates that the total sum of squares of deviations of the observed variable average saccade amplitude is 37719.600, the mean square deviation of different interface layout is 1.600 and 991.103 respectively, dividing 1.600 by 991.103 to get F is 0.002, the corresponding significant level is 0.046, less than the significance level 0.05.

Table 1. The Analysis of one-way ANOVA based on saccade amplitude

	SSD	DOF	Average of SSD	F	Significant level
Inter-group	1.600	1	1.600	0.002	0.046
Intra-group	37718.0	38	991.103		
Sum	37719.6	39			

(SSD = sum of squares of deviations, ANOVA = analysis of variance)

Comparing with the number of fixation points observed by the second group of test subjects, as shown in Fig. 10. It can be concluded from the following diagram that the average number of fixation points of scheme A is 58, the standard deviation is 2.5948, the average number of fixation points of scheme C is 59, the standard deviation is 3.5254, as shown in Table 2. It can be seen that the design of scheme A is better than the design of scheme C.

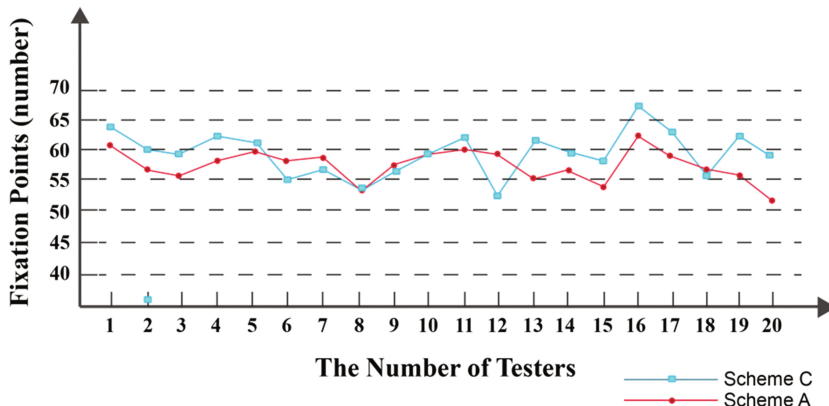


Fig. 10. Fixation points observed by the second group of testers

Table 2. Statistical tables of fixation points

	Scheme	Sample size	Mean	Standard deviation
Number of fixation points	A	20	57.55	2.5948
	C	20	59.4	3.5254

The original hypothesis is that it had no significant effect on the fixation points between scheme A and scheme C, let's suppose that the significant level was 0.05, and the results were shown in Table 3. Based on the above data, we can draw a conclusion that the variance of the sampling error is 0.021, and the mean squared deviations of them are 0.001 and 0.021, dividing 0.001 by 0.021 to get F is 0.048, and the corresponding probability P is 0.341, which is greater than the significance level 0.05, so the hypothesis has no significant effect on the fixation time.

Table 3. The Analysis of one-way ANOVA based on fixation points

	SSD	DOF	Average of SSD	F	Significant level
Inter-group	0.001	1	0.001	0.048	0.341
Intra-group	0.021	38	0.021		
Sum	0.022	39			

2.3.2 Experimental Results and Discussion

From the analysis of saccade trajectory and average saccade amplitude, the interface design of scheme A is more reasonable than scheme B. From the analysis of the fixation points, the shape design of the scheme A is better than the scheme C, that is to say, the interface design of scheme A is more friendly and higher recognition.

However, it can be seen from the One-way ANOVA that the significant number of fixation points in the second group was significantly greater than the level of 0.05, that is we can accept the original hypothesis. There was no significant effect on the number

of fixation points with different shape designs, therefore, it is difficult to evaluate the advantages and disadvantages of the two schemes on shape design. There are some reasons for the problem: sampling error of testing personnel, the arrangement of test sequence has some influence on the experimental results and the error caused by test instrument.

3 Conclusion

In this paper, a systematic and comprehensive study on human cognition and behavior characteristics is presented:

- By using GOMS model to analysis the astronaut's cognitive operation on the human computer interaction system, to extract the design guidance of the man-machine interface, so that the design of the man-machine interface can better play the role of the astronaut in the human-computer interaction system.
- By analyzing the data of the average saccade amplitude and fixation points in the eye movement experiment, we find out rationality and deficiency of scheme. This paper presents a set of analysis and optimization method based on eye-tracking.
- Combining theoretical analysis with experimental verification, a reasonable and systemic design method is proposed for display & control device. It will make a contribution to the design and optimization of other human-machine interaction system.

Acknowledgments. The paper is supported jointly by Science and Technology on Electro-optic Control Laboratory and National Aerospace Science Foundation of China (No. 20165169017), SAST Foundation of China (SAST No. 2016010) and National Natural Science Foundation of China (No. 71471037, 71271053).

References

1. Boney, M.C.: CAPABLE - a computer program to layout controls and panels. *Ergonomics* **20** (3), 297–316 (1977)
2. Dan, M.P.: Using man modeling CAD system and expert systems for ergonomic vehicle interior design. In: Proceedings of the 13th Triennial Congress of the International Ergonomics Association, Tampere, Finland, 29 June–4 July 1997, pp. 80–83 (1997)
3. Fang, W.: Influence of parameter adjustment on the visual effect of the liquid crystal display console. *J. Railway Sci.* **12**, 40–44 (2003)
4. Liu, W.: Situated Cognition in Human-Computer Interaction Theory and Application, pp. 15–68. China science and Technology Press, Beijing (2005)
5. Kieras, D.: GOMS models for task analysis. In: Diaper, D. (ed.) Handbook of Task Analysis for Human-Computer Interaction, pp. 83–116. Lawrence Erlbaum Associates, Mahwah (2003)
6. Carlos, H., Morimoto, M.R.M.M.: Eye gaze tracking techniques for interactive applications. *Comput. Vis. Image Underst.* **98**, 4–24 (2005)

7. Sun, R., Tian, C.: Eye movement analysis technique and application in aviation field. *J. Civ. Aviat. Univ. China* **27**(4), 1–4 (2009)
8. Zhuang, D.: *Theory and Application of Pilot's Attention Allocation*, pp. 31–38. Science Press, Beijing (2013)
9. Yang, X.: Analysis of variance analysis method: one-way ANVOA. *Exp. Sci. Technol.* **11**(1), 23–25 (2013)

The Future Flight Deck

Don Harris^(✉)

Coventry University, Coventry, UK
don.harris@coventry.ac.uk

Abstract. The future commercial flight deck will need to consider the effects of global economic drivers in its design. These issues will considerably alter operating concepts and have a knock-on effect to the human aspects of design and operations. It is argued that ‘user-centered’ design is limited in considering such factors and a more ‘use centered’ design approach is required.

Keywords: Macro-ergonomic drivers · Use centered design · Aviation · Flight deck

1 Introduction

There are many projects in Europe looking at the equipment and functions of the commercial flight deck of future aircraft. The Future Flight Deck project was a major research project, part funded by Innovate UK. This was followed by Open Flight Deck (also funded by Innovate UK). On a European scale, in recent years there have been projects such as the Advanced Cockpit for the Reduction of Stress and Workload (ACROSS) project (see <http://www.across-fp7.eu/>) and REACTOR Reducing Workload Through Efficient Technology and Procedures (an EU Clean Sky 2 programme). The object of such research and development programmes has been to develop new flight deck architectures and their related technologies including displays (Head Up; Head Down and Head Mounted), integrated pilot interfaces, data networks, touchscreen and voice interfaces, haptic interfaces, graphics capabilities and the computer processing on which to implement them. The advanced capabilities will improve the availability of the aircraft by providing the pilot with a fuller picture of the aircraft situation, supporting their decision-making process and optimizing the availability of the aircraft across a range of operational scenarios. Throughout these projects the consideration of Human Factors aspects of the new technology informed design decisions, enabling more radical approaches to flight operations to be evaluated.

Somewhat coincidentally, the Innovate UK-funded Future Flight Deck project commenced almost exactly 20 years after a position paper of the same name was presented, authored by the Flight Operations Group of the Royal Aeronautical Society in conjunction with the Guild of Air Pilots and Air Navigators. In this paper the modern flight decks of the time were subject to analysis and criticism and a view was taken concerning the required developments. In this paper several key areas were discussed:

- The role of the pilot and the development of automation
- Flight deck layout and working environment

- Instrumentation (the transition to multifunction screens on the flight deck)
- Flight Management Systems (FMSs)
- Autopilot and Autothrust (including feedback through the control column or side-sticks)
- New Technology (encompassing things such as civil Head Up Displays – HUDs and Synthetic Vision Systems – SVSs)

While an interesting (and still relevant) set of areas for development it can be seen that the list of topics was very technology-centric. To a certain extent, the current Future Flight Deck program also adopts such a stance, however it does recognize that the design of the next generation of flight decks will be driven by the requirements to support aircraft operation in a Single European Airspace (SESAR)/NextGen (Next Generation) airspace. Flight decks must support associated concepts such as 4-D (four-dimensional) flight planning and zero visibility landing to extend the operational envelop and offer significant fuel savings. Environment and operating concerns are now starting to drive the pilot interface and much as more traditional factors.

The Aerospace Technology Institute (ATI) in the UK is charged by the government to implement the national aerospace technology research strategy by working collaboratively with industry, government and academia (see <http://www.ati.org.uk/about-us/institute/>). The ATI has identified key strategic areas for UK Research and Development but has also defined three key time frames: shorter term development goals (running until 2020 with a target for implementation by 2025); medium term goals to exploit new technologies (for implementation by between 2025–30) and longer term strategic goals stretching from 2030 and beyond (2035+).

These time scales may seem far reaching, but when it is considered that the typical design cycle for a modern commercial aircraft is around seven years and that technologies need to be at around TRL (Technology Readiness Level) nine at the beginning of the development cycle if they are to be incorporated in the final design, suddenly the pressure to develop new technologies becomes evident if the next generation of airliners are not to be out of date the minute that they enter service.

However, the greatest problem may not be in anticipating the technologies required but in anticipating the operational environment/context.

2 Looking Back 20 Years

Two decades ago the first of the UK low-cost carriers, easyJet, has just commenced operations from Luton Airport in the UK using Boeing 737 aircraft. The carrier was based upon the model being used by SouthWest in the United States. Now, 20 years later, much of what was radical in the manner in which the low-cost airlines operated is common practice, even in major carriers. At the same time, navigation practices were still largely based around beacons on the ground and designated airways. Satellite navigation was the exception and not the norm. Air Traffic Control/Air Traffic Management (ATC/ATM) practices across Europe were becoming more harmonized but were still largely dictated by national boundaries.

The military services were also undergoing a period of change, with the introduction of increasing levels of technology coupled with the downsizing of forces. This was starting to result in a shortage of trained pilots emerging from the Air Forces. Around 1998 the global oil price was rising as a result of increasing tensions surrounding Iraq and Iran, and China was emerging as an economic power, entering a period of sustained growth (which would continue). Furthermore, despite these factors the demand for air travel has more than doubled in the last 20 years (between 1990 and 2010 passenger seat kilometers flown worldwide increased from 2,000 billion to over 4,700 billion – [1], with China in particular, showing massive growth in passenger demand, quadrupling in this period.

The point of this brief discussion is simple: the future will not be the same as the present, although what the future holds in the next two decades is difficult to say. However, the aircraft currently being operated are very much the same as those employed by the airlines 20 years ago, albeit with some evolutionary developments.

3 What Does the Future Hold?

The ATI research strategy looks forward, to beyond 2035. Although the shape of the future cannot be predicted, some global drivers (for good or bad) can be anticipated:

- Oil price: There is relationship between oil price, airline operating economics and the demand for air travel. What will happen if the oil price either continues to fall or rises sharply?
- Environmental issues: There will be a continued push for the airline industry to become ‘greener’. Green issues (potentially related to taxes) may also ultimately serve to depress the demand for air transport.
- Capacity: There will be increasing demands on both sector and runway capacity, particularly in the shorter term.
- Cost: To remain competitive and satisfy increasing demand for air travel airlines will strive to contain costs (especially if oil prices or taxation rates begin to rise).
- Emerging markets: South East Asian economies (China, Taiwan, Korea and new players, such as Vietnam) will initially continue to grow but over 20 years also have the potential to slow down or decline. Other markets may continue to grow (India, Brazil) or may suffer the same fate.
- Political instability: Few people could have predicted the rise of so-called Islamic State and the instability in Ukraine, both of which have served to compromise aviation safety.
- Pilot availability: With a decline in the number of trained pilots emerging from the armed forces, coupled with an increasing demand for air travel there may be a chronic shortage of pilots in the next few years. However, if this demand for pilots is satisfied there may subsequently be a glut if global economies slow down.
- New routes: New routes will open, to both smaller airports (more direct routing) or there may be the potential for low cost carriers to start operating on inter-continental routes.

These may all be interesting factors from the perspective of airline operating economics, but what implications do they have for Human Factors in general and the design of future flight decks in particular? It is argued that all of the above have design consequences associated with them which have often been neglected by the Human Factors profession. Furthermore, a subtly different Human Factors design paradigm may be required to satisfy the future requirements of the aviation industry.

4 Design Implications for the Flight Deck of Global Economic Drivers

The future will not be the same as the past, therefore the future flight deck must try and anticipate the effects of macro-economic factors. Human Factors practitioners and researchers need to develop an element of prescience and flexibility in approach. Fortunately, some of the potential knock-on effects of the factors described previously may be anticipated, to a degree.

4.1 Pilot Shortage

There are already signs of a rapidly increasing shortage of commercial airline pilots. Boeing estimate that between 2015–34 95,000 commercial pilots will be required in North America alone versus a potential supply of only 64,000 in this period. This is traditionally seen as a pilot recruitment and training issue. However, there are fundamental design considerations that are driven and can potentially help to alleviate this anticipated shortage. There will be a tendency over the next 20 years for experience across the flight deck to decrease, with many more low-hours captains being teamed with an increasing number of low-hours First Officers.

Good human-centred design places the target audience description at the centre of the design process. To accommodate the change in the end user-group the complexity of flight deck interfaces will need to be reduced, which will also have the additional bonus of decreasing training time. It will also be argued that the Human Factors design approach adopted will have to change.

One option to address the potential shortage of pilots and help to further reduce costs is the development of single pilot commercial aircraft. The trend in flight deck design over the past half century has been one of progressive ‘de-crewing’. Aircraft manufacturers and avionics systems suppliers (e.g. Embraer and Honeywell – [2]) are developing the advanced technology for such aircraft, centred upon the development of Intelligent Knowledge-Based Systems and adaptive automation. An alternative design approach proposed initially by Harris [3] uses a distributed systems-based design philosophy utilizing a great deal of extant technology derived from single seater military aircraft and UASs technology. In this case the control and crewing of the aircraft is distributed in real time across both the flight deck and ground stations [4].

4.2 Low Cost Carriers

Low cost carrier concepts of operation are now the norm, even in larger airlines. However, the aircraft being operated have never been designed (from a flight deck viewpoint) for intensive, short(er) range operations with rapid turnaround times on the gate. Many autopilot modes are not necessary for such operations. Airline personnel costs vary between about 11% of operating costs to nearly 25%, depending upon aircraft type, sector length and how much activity is outsourced [5, 6]. Annual accounts from a typical low-cost operator suggest that even for a larger airliner, the crew represent nearly 13% of operating costs (excluding fuel and propulsion – [6]).

Gate time is charged by the minute. Reducing time on the gate can result in considerable cost savings. Pilot training is another major overhead. Reduction of initial and recurrent training time will also both reduce costs and increase pilot availability.

Many low cost carriers also operate into smaller, secondary airports around major cities. These are often less well equipped than the major hubs. While this can serve to reduce delays, in bad weather such airports can be difficult to navigate and have considerable restrictions on arrivals and departures. Increasing aircraft availability by allowing poor visibility approaches, landings and take-offs, in conjunction with aiding low visibility taxiing will considerably reduce operating costs resulting from delays. However, these capabilities must be present in the aircraft and not dependent upon ground infrastructure.

4.3 Green Issues

There are undoubtedly going to be continued demands to make aviation more environmentally friendly, reducing its carbon footprint (e.g. see the major initiative undertaken by the European Union in its Clean Sky and Clean Sky 2 research programs). These have both direct and indirect Human Factors considerations for the future flight deck.

Direct considerations include initiative such as flying direct routings and optimizing climb and descents profiles, both likely to be undertaken in increasingly congested airspace. This will require increased levels of automated assistance for flight planning and execution, and the display of complex departure and arrival procedures. The three-dimensional display of complex routings (especially departure and arrival routes) and 3D depiction of weather – especially winds will aid in this respect. The display of traffic/collision alert information might be useful, especially as under SESAR/NEXTGEN, self-assured separation is a key concept.

Indirect Human Factors considerations will be related to the design of the aircraft. There are concepts being developed for jet transport aircraft which will have high aspect ratio wings, allowing higher, more economical flight but will at the same time considerably reduce the cruising speed of the aircraft. This obviously has implications for pilot fatigue. Research into the operation of long-haul aircraft during the cruise phase using just a single member of flight was undertaken in the Advanced Cockpit for the Reduction of Stress and Workload (ACROSS) project (see <http://www.across-fp7.eu/>).

This will also reduce the operating costs associated with the need to carry a third pilot during ultra-long haul operations, specifically for the approach and landing phase.

4.4 Political Instability

Political instability may have several effects depending upon its nature and location. Oil process may increase considerably (further enhancing the pressures of on operating costs). However, there are also increasing threats to aircraft from the ground (e.g. ground to air missiles) from insurgent groups engaged in local conflicts. There may be requirements for in-flight re-routing, and updating of threat information (much as the updates of tactical information received by military aircraft).

Oil is traded in dollars. Brexit has had the effect of increasing fuel prices in the UK as the exchange rate between the pound sterling and US dollar deteriorates.

4.5 Culture and Emerging Markets in South East Asia

Boeing market analysis suggests that the Asia Pacific Region will require substantially more aircraft over the next two decades making it the largest airline market in the world (2012 fleet - 5,090 aircraft; projected 2032 fleet - 14,750 aircraft). Thirty-five percent of the world fleet will be domiciled in this region.

However, cultural issues on the flight deck run deeper than issues in Crew Resource Management [7]. The operating philosophy on civil flight decks is based upon two pilots cross-monitoring the other's actions. The system of 'monitor and cross-monitor' and 'challenge and response' is predicated upon the assumption that crew will speak up and alert each other to irregularities and errors but this is implicitly based upon a Western cultural assumption. There are also fundamental differences in the mental models of people in European/North American and Chinese cultures. Westerners adopt a function-oriented mental model connected to a task-oriented operating concept (where specific actions are performed to achieve well-defined results) resulting in a preference for a sequential approach to undertaking tasks. The Chinese preference is for a more holistic integrated, thematic approach hence the task-oriented operating concept contradicts their preferred method of working. A multi-configurable flight deck interface may provide manufacturers with a competitive advantage in some markets.

4.6 Cost

Cost drives everything. Pilot shortages increase the market price of pilots, increasing cost. Longer routes to avoid areas of political instability increase costs. Aircraft that fly more slowly to reduce fuel burn will also decrease utilization, potentially offsetting the decrease in operating costs attributable to reduced fuel burn. Green issues may restrict routing options and departure/arrival times.

5 User vs. Use Centered Design

The Human Factors profession has traditionally been dominated by a ‘user-centered design’ approach, and the aerospace industry is no exception. Emphasis has been placed on primarily on ensuring that equipment and procedures on the flight deck are commensurate with the skills, knowledge and abilities of the end users (i.e. pilots).

‘Use-centered’ design adopts a more socio-technical system oriented stance, which includes the work domain as a third component of the system [8]. The work domain provides further constraints on the work system. The envisaged global economic drivers fall into this category. It is argued that the macro-global trends may require a change in the Human Factors approach to be adopted. While the human operator (pilot) is still placed at the center of the flight deck design process, placing emphasis on ‘use’ as opposed to ‘user’ requires a change to a more problem-centered focus for flight deck design.

This can be expressed another way: if we can’t get enough of the current type of pilot we will need to design aircraft for a different type of pilot that we can get enough of.

6 Flight Deck Design: Re-visiting the Past

Many of the complexities on the flight deck result from the fact that they are full of outdated systems, a legacy of the old technologies that were implemented on the flight deck of aircraft over three decades ago. To make change even more difficult, these are now mandated in the international airworthiness requirements (for example the US Federal Aviation Requirements and the EU Certification Specifications for large aircraft). These complexities partially dictate the skills and knowledge required of the modern pilot; they serve to limit the number of potential candidates for pilot training; increase the complexity of training, and hence increase the market price and availability of trained, experience aircrew.

What is more, these complex, old-fashioned legacy systems are no longer required. For example, with satellite-based technology a pilot can know exactly how high they are. Barometric altimetry, which used the changes in air pressure with altitude, required at least three different definitions of ‘height’ when flying: height relative to the local terrain; height above mean sea level (altitude) and Flight Level (a notional altitude where everyone set their altimeters to be calibrated to 1013 mb). These demanded a series of transition altitudes around airports and in airspace. Speed was a similar nightmare. Ground speed is speed over the face of the Earth; true airspeed is the relative velocity between the aircraft and the surrounding air mass, which is effected by headwind or tailwind components; indicated airspeed is the speed on the airspeed indicator on the flight deck, but this is also effected by altitude and temperature (density), so becomes increasingly divergent from ground speed as the aircraft climbs. Mach is aircraft velocity relative to the local speed of sound... Keeping the aircraft flying is a problem in airspeed; navigation is a problem in ground speed. Heading refers to compass heading, which is relative to magnetic North (not true North) and indicates which way the aircraft is pointing – but not which way it is travelling because of the effects of cross winds. Track refers to the path of the aircraft across the face of the

Earth. With regard to vertical navigation (VNAV) no current aircraft control modes relate directly to the three-dimensional navigation solution as they are aircraft referenced, not Earth referenced (with the exception of approach mode which is referenced to a fixed point on the planet – the runway). In other words, they do not directly support the pilots' primary task. There are several aircraft-referenced modes (vertical speed; flight path angle; climb-to-speed). The pilot cannot, however, follow a prescribed track in three-dimensional space (which is actually what Air Traffic Control requires).

With modern inertial, Doppler radar and or satellite navigation systems none of these distinctions (and separate parts of instrumentation) are now required; speed is speed; height is height and heading is heading. All of these issues are easily solved with on-board computing technology and digital fly-by-wire systems. The technology is available to make flying an aircraft manually almost as simple as driving a car in three-dimensional space (well, to an extent)! This now changes the nature of the potential user (pilot). By changing from a user-centered design perspective driven by current pilot aptitudes to a use-centered design perspective, where the method of operating the aircraft is driven by the environment (work domain) and the technology available, a whole new potential population from which pilots can be drawn is liberated.

However, the flight deck design regulations are now inhibiting change [9] and hence also constraining other aspects of the sociotechnical system of airline operations (e.g. pilot selection and training).

This will not be a popular perspective. Nevertheless, it serves to illustrate that the Human Factors profession should not always be driven by user centered design or 'fitting the job to the person' [10]. Perhaps a new perspective - 'fitting the new person to the new job' is required?

References

1. IATA: World Air Travel and World Air Freight Carried 1950–2010. <http://people.hofstra.edu/geotrans/eng/ch3en/conc3en/evolairtransport.html>. Accessed 3 July 2010
2. Keinrath, C., Vašek, J., Dorneich, M.: A cognitive adaptive man-machine Interface for future Flight Decks. In: Droog, A., Heese, M. (eds.) Performance, Safety and Well-being in Aviation Proceedings of the 29th Conference of the European Association for Aviation Psychology, Budapest, Hungary, 20–24 September 2010. European Association of Aviation Psychology (2010)
3. Harris, D.: A human-centred design agenda for the development of a single crew operated commercial aircraft. *Aircr. Eng. Aerosp. Technol.* **79**(5), 518–526 (2007)
4. Harris, D., Stanton, N.A., Starr, A.: Spot the difference: operational event sequence diagrams as a formal method for work allocation in the development of single pilot operations for commercial aircraft. *Ergonomics* **58**(11), 1773–1791 (2015)
5. Ryanair: Full Year Results (2009). http://www.ryanair.com/doc/investor/2009/q4_2009_doc.pdf. Accessed 17 November 2017
6. easyJet plc: Annual report and accounts 2013 (2013). <http://corporate.easyjet.com/~media/Files/E/Easyjet-Plc-V2/pdf/investors/result-center-investor/annual-report-2013.pdf>. Accessed 27 June 2014

7. Harris, D., Li, W.-C.: Cockpit design and cross-cultural issues underlying failures in crew resource management. *Aviat. Space Environ. Med.* **79**(5), 537–538 (2008)
8. Flach, J.M., Dominguez, C.O.: Use-centered design: integrating the user, instrument, and goal. *Ergon. Des.* **3**(3), 19–24 (1995)
9. Harris, D.: *Human Performance on the Flight Deck*. Ashgate, Aldershot (2011)
10. Kroemer, K., Grandjean, E.: *Fitting the Task to the Human*, 5th edn. Taylor and Francis, London (1997)

Automated Online Determination of Pilot Activity Under Uncertainty by Using Evidential Reasoning

Fabian Honecker^(✉) and Axel Schulte

Institute of Flight Systems (IFS), Bundeswehr University Munich,
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
{fabian.honecker, axel.schulte}@unibw.de

Abstract. The objective of a workload-adaptive associate system is to support human pilots in critical workload situations to avoid excessive demands on their mental capacity. Since workload strongly depends on the current activity of the pilots, i.e. the tasks the pilots are performing in a specific task situation, a key feature of an adaptive associate system is to determine the activity of the pilots online in real-time.

This contribution presents a method for determining the activity of helicopter pilots automatically in an uncertain and complex environment like military manned-unmanned teaming missions (MUM-T), where multiple unmanned aerial vehicles (UAVs) are commanded from the cockpit of a manned helicopter. We use a pilot observation system with different measurement sensors to collect evidences and apply an evidential reasoning algorithm to draw conclusions on the actual activity of the pilots. Our method is based on a simplified version of Dempster-Shafer theory and is capable of collecting and combining even contradictory evidences.

By providing a means of implicit deliberative communication, this method lays a foundation for improving human-machine team performance in complex task situations. The implementation of this method in a helicopter mission simulator is explained in detail.

Keywords: Activity determination · Adaptive associate systems · Dempster-Shafer theory · Mental workload · Pilot monitoring

1 Introduction and Problem Description

The domain of military helicopter missions is characterized by complexity and uncertainty, especially if the pilots command several unmanned aerial vehicles (UAVs) from the cockpit of a manned helicopter (manned-unmanned teaming MUM-T) [1, 2]. During their mission, the pilots must perform many cognitive tasks that require different amounts of mental resources. While processing those tasks, the workload of the pilots can widely change with different task situations. Especially in situations, in which the pilots are overloaded, decrements in human performance might occur. These decrements in human performance can have a negative impact on the total performance of a helicopter mission.

To counteract performance decrements in high workload situations, automation is often used to reduce mental workload (MWL) of the operator of a technical system. But

the use of automation is not carefree and human factors problems play the central role. Among them are deskilling of the operator [3], boredom, complacency effects and clumsiness of the automation [4, 5], mode errors and mode confusion [6], loss of situation awareness [7], as well as design factors like complexity, brittleness, opacity and literalism [8].

To tackle those human factors problems in the domain of aviation, cognitive automation and cognitive associate systems have been developed over the years to support the human pilots [9–13]. In order to keep the pilots in the loop and to not disturb the work process, this support should adapt to the tasks the pilot is currently conducting. Furthermore, the goal of adaptive automation is to keep the operators' workload at an average level to maintain human performance in extreme workload situations [14, 15]. Workload-adaptive associate systems use adaptive automation to support the human operators.

According to [16], MWL is a multidimensional construct, which is determined by characteristics of the task, of the operator and, to a degree, the environmental context. Our current approach towards engineering a workload-adaptive associate system is given by [17, 18]. It is based on a context-rich description of MWL. A basic assumption in this approach is, that the MWL of the operators follows qualitatively the current task load induced by the task situation [19]. Therefore, a task-centered design has been proposed. With the application in an associate system in mind, in this concept the construct MWL is operationalized context-rich in the form of a plan (the tasks an operator has to do), the current activity, demands on mental resources associated with this activity and observable behavior patterns. The necessary knowledge is stored in a common task model. In order to derive mental workload, the mission has to be planned [20], the current activity determined, demands on mental resources estimated [13] and behavior patterns analyzed [21]. A requirement for determining MWL in a dynamic mission and specific task situation is a reliable determination of the current activity of a human operator.

In this task-centered approach, the human operators are communicating with artificial team members. According to [22], anticipatory information sharing by implicit deliberative communication can improve team performance in complex task situations. Our task-centered method provides a means of implicit, non-verbal deliberative communication, since the human pilots inform the artificial crew member implicitly about their tasks and therefore their goals. Our method for activity determination lays a foundation for a better performance in mixed human and machine teams like the manned-unmanned teaming of a helicopter and UAVs. The following sections give deeper insights, how activity determination can be designed and implemented by using evidential reasoning.

2 Method

2.1 Requirements for Activity Determination

Activity determination can be regarded as a classification problem. If the current activity of a human operator is described by a set of tasks contained in a task model, the method must determine for every task, if the operator is currently conducting this task or not. For a successful and robust determination of the activity, different requirements must be met:

Activity determination must work in an environment, which is characterized by uncertainty and ignorance. Since there is no direct link into the human brain, measurement sensors like buttons, speech detectors or gaze trackers must be used to draw conclusions regarding the underlying cognitive processes. These sensors introduce uncertainty because each measurement has an individual measuring inaccuracy. A consequence is that generated hypotheses based on those measurements have limited reliability. Ignorance results from missing knowledge about model parameters and the environment. If a sensor is temporarily out of order, all measurements of this sensor are completely unreliably. This is for example the case for a gaze tracking system. Loss of gaze tracking can occur, if the test person is blinking or covering a camera. Not only the measurement equipment, but also the human behavior can be faulty. An easy example would be if a pilot forgets to set some system parameters like radio frequencies. Therefore, the algorithms must be able to deal with uncertainty and ignorance and weight the resulting hypotheses of the sensors according to their measurement accuracy. Since activity determination must work automatically in real-time as part of an associate system, these algorithms must also be fast enough. The cognitive processes inside the human brain run on a time scale of tens to hundreds of milliseconds [23]. Therefore, activity determination should also work on this time scale.

Methods that meet those requirements and which seem promising for activity determination are probability theory, Dempster-Shafer theory (DST) [24, 25] and certainty factors [26]. Because of its simplicity and mathematical foundation, probability theory, and especially Bayesian Networks [27], are suited for solving classification problems [28]. For example, Naïve Bayesian Classifiers can be used to filter junk email [29, 30]. Bayesian Networks model the reasoning process in the causal direction and use Bayes' Rule to solve diagnostic problems. The disadvantage of using probability theory is that all model parameters (a-priori and conditional probabilities) must be determined from statistical analysis prior using the model. Certainty factors are a heuristic approach, where the parameters are obtained from expert knowledge. Wittig has used certainty factors for determining pilot intent [31]. DST is also a theory based on a mathematical framework. From a certain point of view, it can be regarded as an extension of probability theory. In contrast to probability theory, DST distinguishes between uncertainty and ignorance. Considering ignorance enables an associate system to be aware of its own lack of knowledge, which might be useful when it comes to critical decisions. In those situations, it is usually better if the human operator decides what to do, rather than a system with uncomplete knowledge. DST is a method for evidential reasoning and extends probability theory by providing a rule of combination for diagnostic problems. The difficulty in DST is the interpretation of the parameters and how they are acquired. Beside certainty factors, DST has also be applied in the MYCIN expert system for computer based medical consultations [32].

Because of providing a rule how to combine evidences, the ability to consider both uncertainty and ignorance and its mathematical foundation, we suggest to take DST for pilot's activity determination. We are using a simplified version of DST based evidential reasoning because the drawback of the full theory is an exponential complexity in calculation and therefore difficult to implement in real-time [32].

2.2 Evidential Reasoning

Since the execution of tasks, especially cognitive tasks, cannot be directly observed, we are using an indirect method for classifying, whether a task is currently being executed by a pilot or not. The used evidential reasoning method is similar to the way a human observer would take (see Fig. 1). During the observation of the pilots, the technical system, and the environment, many observable facts are collected. Each of those facts proposes a hypothesis, whether an operator is currently executing a certain task or not. These single evidences are weighted according to their plausibility and combined into one single resulting hypothesis.

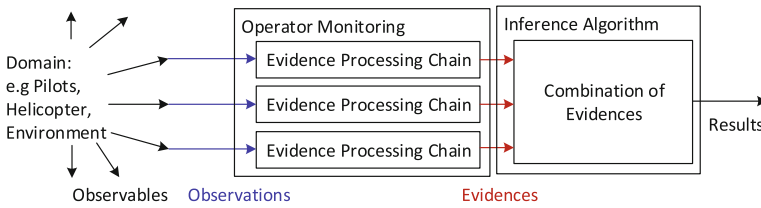


Fig. 1. Process of evidential reasoning in the domain of helicopter missions.

The underlying reasoning model is described in the Sects. 2.3 and 2.4. The processing chain from collecting observations to generate evidences is explained in Sect. 2.5. The inference algorithm for combining evidences to draw conclusions on the actual activity is explained in Sect. 2.6.

2.3 Representing Uncertainty and Ignorance

In probability theory, the strength of belief in a hypothesis X is represented by the scalar probability $P(X)$. $P(X)$ describes the belief in this hypothesis and $1 - P(X)$ the doubt. In contrast to probability theory, we represent the strength of belief in a hypothesis X by a belief triplet according to [33]:

$$Q(X) = (p, q, r) \tag{1}$$

This triplet is a normalized distribution:

$$p + q + r = 1 \tag{2}$$

The belief quantities p, q and r can have continuous values in the range from 0 to 1. The quantity p is called belief and signifies to which extent the hypothesis is supported, q the doubt, to which extent the hypothesis is rejected, and r the remaining ignorance. There are different interpretations of belief values. According to the interpretation of Dempster [24], the ignorance r describes uncertainty in quantifying probabilities, p is a lower probability value and $1 - q$ an upper probability value (plausibility), where the actual probability lies somewhere in between. Other interpretations do not try to relate belief values directly to probabilities. Shortliffe, for example, interprets belief values in certainty factor theory as

an increase in information and gives the following example: “I don’t know what the probability is that all ravens are black, but I do know that every time you show me an additional black raven my belief is increased by X that all ravens are black” [26].

Similar to conditional probability tables, conditional belief distributions can be defined. These distributions are described in our method by matrixes:

$$M_S(B|A) = \begin{pmatrix} p_t & p_f & 0 \\ q_t & q_f & 0 \\ r_t & r_f & 1 \end{pmatrix} \tag{3}$$

In this definition, the matrix components significate the following:

- p_t : Belief, that $B = true$ if $A = true$.
- q_t : Belief, that $B = false$ if $A = true$.
- r_t : Belief, that $B = unknown$ if $A = true$.
- p_f : Belief, that $B = true$ if $A = false$.
- q_f : Belief, that $B = false$ if $A = false$.
- r_f : Belief, that $B = unknown$ if $A = false$.

2.4 State Space Model

We are using a separate classification model for every possible task that can be conducted by a test person. The state space model for a single task, which classifies whether it is part of the current activity or not, is depicted in Fig. 2.

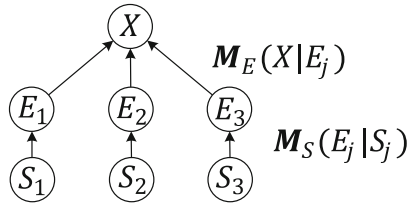


Fig. 2. State space model.

The state space model is a structured representation of the dependencies between a task, evidences and observations. It consists of nodes and edges and is similar to a Bayesian Network. The nodes symbolize state variables described by belief triplets (1). The root node is the state variable for the examined task X . Several evidence variables E_1, \dots, E_m are linked to the task X . Since each evidence is based on a hypothesis generated by a measurement sensor, each evidence variable E_j is connected to a sensor variable (observation) S_j . The edges in the state space model symbolize reasoning models described by conditional belief distributions (3). The arrows indicate the direction of belief propagation for diagnostic reasoning. There are two types of models: Sensor models and evidence models.

A sensor model describes the accuracy or reliability of a measurement sensor. Beside the general sensor model $M_S(E_j|S_j)$ according to (3), we define some special types of a sensor model:

$$M_S(E_j|S_j) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{Perfect Sensor} \quad (4)$$

$$M_S(E_j|S_j) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{Unknown sensor} \quad (5)$$

$$M_S(E_j|S_j) = \begin{pmatrix} P_t & P_f & 0 \\ 1 - P_t & 1 - P_f & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{Probabilistic sensor} \quad (6)$$

$$M_S(E_j|S_j) = \begin{pmatrix} Z & 0 & 0 \\ 0 & Z & 0 \\ 1 - Z & 1 - Z & 1 \end{pmatrix} \quad \text{Sensor with scalar reliability} \quad (7)$$

For a perfect sensor, both, uncertainty and ignorance are 0. An unknown sensor model is the contrary of a perfect sensor model and will always result in ignorance 1 for any input. For a Bayesian definition of the sensor model with no ignorance, the probabilistic model is used and for representing reliability and the amount of knowledge, a simple sensor model with a single scalar value Z might be used.

The evidence model describes the strength of an evidence for task execution under the assumption, that there is no measurement error by the sensor (perfect sensor model). In contrast to the sensor model, the evidence model contains human factors like uncertainty in human behavior. A general model of an evidence can be written in matrix form $M_E(X|E_j)$ according to Eq. (3). Beside a perfect, unknown and probabilistic evidence (compare with sensor models), we consider two special types of evidence models:

$$M_E(X|E_j) = \begin{pmatrix} p & 0 & 0 \\ 0 & 0 & 0 \\ 1 - p & 1 & 1 \end{pmatrix} \quad \text{Supporting evidence (belief)} \quad (8)$$

$$M_E(X|E_j) = \begin{pmatrix} 0 & 0 & 0 \\ q & 0 & 0 \\ 1 - q & 1 & 1 \end{pmatrix} \quad \text{Rejecting evidence (doubt)} \quad (9)$$

A supporting or belief evidence, increases the strength of belief that the operator is currently performing a task, whereas a rejecting or doubt evidence increases the strength of doubt in that hypothesis.

2.5 Evidence Processing Chain

For each evidence contained in the state space model, a processing chain is used to calculate its belief values and therefore evidential strength (Fig. 3).

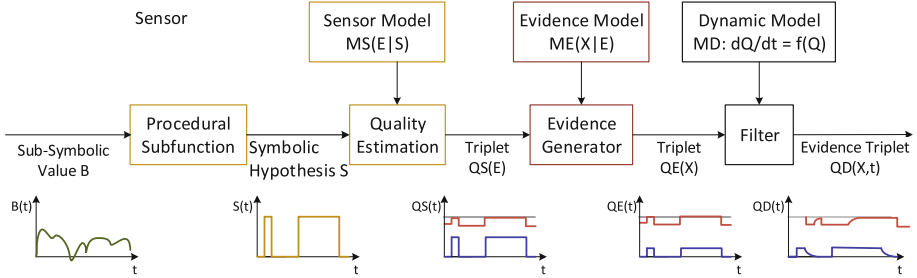


Fig. 3. Processing chain for evidences.

The raw information about the test persons and their environment is given in an inhomogeneous and sub-symbolic form. In the first step (procedural sub-functions), symbolic hypotheses S (observations) are derived from sub-symbolic values B . To be compatible with classical logic and Bayesian models, each observation is described by a state variable (see Sect. 2.4), which can be either *true*, *false*, or *unknown* if the sensor is broken or temporarily out of order. This value is then transformed into a belief triplet:

$$\mathbf{Q}_S(S_j) = \begin{cases} (1, 0, 0)^T & \text{if } S_j = \text{true} \\ (0, 1, 0)^T & \text{if } S_j = \text{false} \\ (0, 0, 1)^T & \text{if } S_j = \text{unknown} \end{cases} \quad (10)$$

The reliability of this hypothesis is calculated as a matrix-vector multiplication of the sensor model with this belief triplet:

$$\mathbf{Q}_S(E_j) = \mathbf{M}_S(E_j|S_j)\mathbf{Q}_S(S_j) \quad (11)$$

Then the strength of a single evidence for a given point in time is derived from a forward propagation of the belief values similar to [34, 35]. For this calculation, the evidence model is used:

$$\mathbf{Q}_{E_j}(X) = \mathbf{M}_E(X|E_j)\mathbf{Q}_S(E_j) \quad (12)$$

In a time-dependent environment, it is not sufficient to describe evidences only for a given point in time. A dynamic model describes, how long an observed evidence is valid and is indispensable if short events like button presses are used as evidences. A general dynamic model can be defined as an initial value problem, i.e. a differential equation with initial conditions:

$$\mathbf{M}_D: \begin{aligned} \frac{d}{dt} \mathbf{Q}_{D_j}(X, t) &= f(\mathbf{Q}_{D_j}(X, t)) \\ \mathbf{Q}_{D_j}(X, 0) &= \mathbf{Q}_{E_j}(X) \end{aligned} \tag{13}$$

We suggest a simple dynamic model, where the belief p and doubt q are decaying after the evidence has been observed. The rate of decay is assumed to be proportional to the current belief values:

$$\begin{aligned} \frac{dp(t)}{dt} &= \lambda p(t) \rightarrow dp = \lambda p dt \\ \frac{dq(t)}{dt} &= \lambda q(t) \rightarrow dq = \lambda q dt \end{aligned} \tag{14}$$

Solving this differential equation leads to an exponential decay over time:

$$\begin{aligned} p_{D_j}(X, t) &= p_{E_j}(X) \exp(-\lambda t) \\ q_{D_j}(X, t) &= q_{E_j}(X) \exp(-\lambda t) \\ r_{D_j}(X, t) &= 1 - p_{D_j}(X, t) - q_{D_j}(X, t) \end{aligned} \tag{15}$$

This model imitates the intuitive observation, that knowledge is getting lost over time and ignorance increases. The decay parameter λ of this model can be expressed as a half-life value, which indicates the time, after which the belief is half of the originally observed one:

$$t_{\frac{1}{2}} = \frac{\ln(2)}{\lambda} \tag{16}$$

2.6 Combination of Evidences

After generating the evidences, in favor for or against the execution of a single task, all evidences are combined into a single resulting hypothesis. This is done by using a rule of combination, which is derived from Dempster’s rule of combination for a binary frame of discernment. With this rule, two evidences are combined by:

$$\mathbf{Q}_1(X) \oplus \mathbf{Q}_2(X) := \mathbf{Q}_{12}(X) = \begin{pmatrix} p_{12}(X) \\ q_{12}(X) \\ r_{12}(X) \end{pmatrix} \tag{17}$$

$$\begin{aligned} p_{12}(X) &= \frac{p_1 p_2 + p_1 r_2 + r_1 p_2}{1 - (p_1 q_2 + q_1 p_2)} \\ q_{12}(X) &= \frac{q_1 q_2 + q_1 r_2 + r_1 q_2}{1 - (p_1 q_2 + q_1 p_2)} \\ r_{12}(X) &= 1 - p_{12}(X) - q_{12}(X) \end{aligned}$$

All evidences resulting from the processing chain are combined by using this rule of combination iteratively:

$$\mathbf{Q}_{total}(X, t) = \mathbf{Q}_{D1}(X, t) \oplus \mathbf{Q}_{D2}(X, t) \oplus \dots \oplus \mathbf{Q}_{Dm}(X, t) \quad (18)$$

The order of the evidences does not matter since the combination rule given above is both associative and commutative. Tasks, for which the resulting triplets \mathbf{Q}_{total} have high belief and low doubt values, are considered as part of the current activity of a human operator.

3 Implementation in a Helicopter Mission Simulator

3.1 System Overview

The described method for activity determination is implemented in the helicopter mission simulator of the Institute of Flight Systems at the Bundeswehr University in Munich. It is part of a situation and workload-adaptive pilot associate system for MUM-T helicopter missions. Figure 4 shows an overview of the system for activity determination.

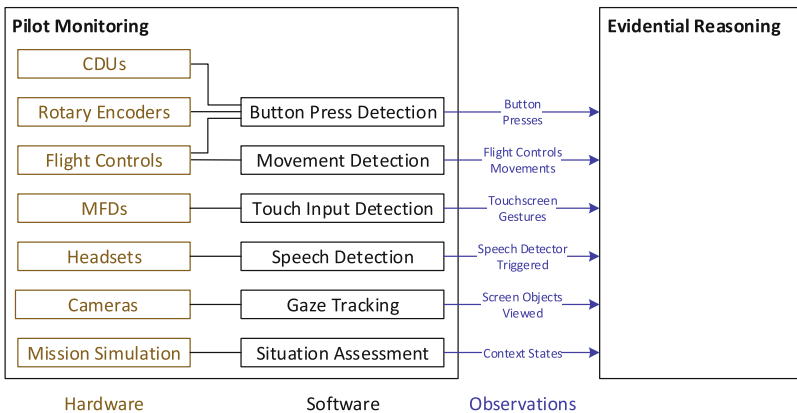


Fig. 4. System overview of the activity determination.

Firstly, observations are created during pilot monitoring, then the actual activity is inferred during evidential reasoning.

As hardware part of the sensors, different input devices are used. Input devices are the command and display units (CDUs), rotary encoders, flight controls, touch-sensitive multi-function displays (MFDs) [36], microphones in the headsets of the pilots and cameras for a gaze tracking system. Furthermore, states of the mission simulations are considered to capture the task context. For each type of these hardware sensors, we calculate observations consisting of symbolic detection values and quality measures to estimate sensor reliability in different software modules. These modules implement the low-level, procedural signal processing sub-functions of the processing chain described

in Sect. 2.5. We explain the process of pilot monitoring for each different sensor type below (Sects. 3.2–3.8).

The inference algorithm is implemented in C++ computer code in the program PAD. The program PAD uses a separate thread for every pilot and does calculations in parallel. The observations from different sensors may have different signal propagation delays. Therefore, in the first step, those delays are corrected by using the actual times of the measurements. This is important to maintain causality and prevent reasoning problems, if later created signals arrive earlier at the activity determination. Latencies in the signals result from inter-process communications between the sensors and the PAD program. Then, the sensor model $M_S(E_j|S_j)$ for a scalar reliability value Z (7) is used to derive the observation triplet for every evidence $Q(E_j)$. The processing chain and the rule of combination are implemented as described in Sects. 2.5 and 2.6. Figure 5 shows an excerpt from the resulting belief distributions.

Right Pilot					
Detected Activity/CommunicateIntern and FlyTransitFriendManual					
	Task	Filtered	Belief	Doubt	Ignorance
1	CommunicateIntern		0.97	0.03	0.00
2	FlyTransitFriendManual		0.95	0.05	0.00
3	CheckRadio		0.20	0.00	0.80
4	CommunicateATC		0.00	1.00	0.00

Fig. 5. Example results for activity determination, screenshot of program PAD. Here, the pilot is flying a manual transit flight and communicating via the intercom at the same time.

The final step is to extract the actual activity of the pilots from those distributions. That means a classification decision must be made. Therefore, the belief triplet must be interpreted. We consider a task as part of the activity, if the belief is the greatest value:

$$p > q + r \tag{19}$$

Since the belief triplet is normalized (2), this criterion is equal to

$$p > 0.5. \tag{20}$$

In practice, taking just the threshold value might result in oscillations and causality problems for very short events (e.g. button presses and state changes at the same time). Therefore, we add a last filtering step. This step is done by integrating the belief over a short time. To select the actual tasks of the activity, an impact value is calculated while the belief is greater than 0.5:

$$I = \int_{\text{while } p(t) > 0.5} p(t) dt \tag{21}$$

If the impact of the evidence is high enough (in our simulator a threshold value of $I = 0.4$ is used), this task is classified as part of the current activity. The integration is stopped if the current belief drops below 0.5. The drawback of this last filter step is an additional time delay of the detection decision. The integration time depends on the strength of belief. The higher p , the shorter the integration time and the faster the decision.

3.2 General Rule for Calculating Sensor Reliability for a Continuous Value

An observable is a measurable quantity in the simulator. Most of the sub-symbolic observables are continuous, time-dependent signals. If an observable is observed during the process of activity determination (i.e. a detector has been triggered), we call it observation. In the implementation, we are using the scalar reliability sensor model (7). Therefore, every observation is described as symbolic hypothesis S along with a scalar reliability or quality value Z .

For generating a symbolic value from a sub-symbolic, continuous quantity x in the procedural sub-functions (Sect. 2.5), we are using a general threshold criterion to distinguish if a detector has been triggered or not:

$$S(x) = \begin{cases} true & \text{if } x \geq x_d \\ false & \text{if } x < x_d \\ unknown & \text{if sensor out of order} \end{cases} \quad (22)$$

The scalar reliability value is calculated as linear distance to the threshold x_d :

$$Z(x) = \begin{cases} \frac{x - x_d}{x_{max} - x_d} & \text{if } x \geq x_d \\ [1mm] \frac{x - x_d}{x_{min} - x_d} & \text{if } x < x_d \\ [1mm] 0 & \text{if sensor out of order} \end{cases} \quad (23)$$

A perfect measurement corresponds to the reliability $Z = 1$ and a totally unreliable measurement corresponds to the reliability $Z = 0$. For practical purposes, the reliability value is bounded, i.e. values below zero are set to 0 and values above 1 are set to 1. Beside the discrimination threshold x_d , this model requires a lower limit x_{min} and an upper limit x_{max} as parameters, which are defined below for every sensor type.

3.3 Button Press Detection

In the cockpit, there are many buttons the pilots can press. Examples are buttons of the CDUs or buttons on the grip of the flight controls (Fig. 6). Buttons can either represent system states (e.g. radio button) or short events, if they are pressed (e.g. line selection keys on the CDU).

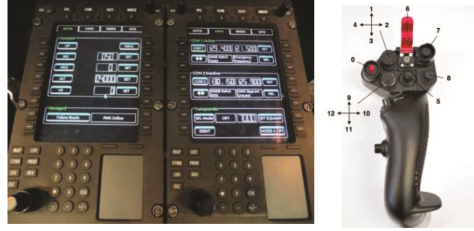


Fig. 6. CDUs and flight controls with buttons [37].

Buttons in aviation must be extremely reliable. Even buttons in the simulator are highly reliable and we do not expect that they fail or deliver inaccurate results during a typical mission of less than two hours. Therefore, a button press is described by a perfect sensor model (4), which is equal to a scalar reliability model (7) with $Z = 1$:

$$S = \begin{cases} \text{true} & \text{if button pressed} \\ \text{false} & \text{if button not pressed} \end{cases} \quad (24)$$

$$Z = 1 \quad (25)$$

If the state of a button can take more than two discrete values, which can be the case for switches and rotary buttons, a binary model is generated for every possible value.

3.4 Movement Detection of the Flight Controls

For steering the simulated helicopter, a control load system of Reiser Simulation and Training GmbH has been integrated into the simulator (Fig. 7). The system consists of a cyclic stick, which controls the pitch and roll movement of the helicopter, pedals which control the yaw movement and a collective lever which controls the collective pitch of the rotor blades. The flight controls of both pilots are electrically coupled so that they are physically performing the same movements in parallel.



Fig. 7. Flight controls of the helicopter simulator: collective, cyclic and pedals.

The movement detection of the flight controls is based on measuring the rate $r(t)$ (i.e. the time derivative) of the flight controls signal $x(t)$ of each axis:

$$r(t) = \left| \frac{dx(t)}{dt} \right| \quad (26)$$

If one of the rates for the different axes is greater than the threshold r_d , the stick is considered as moved by the pilot:

$$S(r) = \begin{cases} true & \text{if } r \geq r_d \\ false & \text{if } r < r_d \\ unknown & \text{if sensor out of order} \end{cases} \quad (27)$$

For estimating the detection quality, the lower limit in Eq. (23) is no movement at all ($x_{\min} = 0$). The upper limit x_{\max} is determined during the calibration for a fast movement of the flight controls. The reliability value depends on the rate of movement:

$$Z(r) = \begin{cases} \frac{r - r_d}{r_{\max} - r_d} & \text{if } r \geq r_d \\ [2mm] \frac{(r_d - r)}{r_d} & \text{if } r < r_d \\ [1mm] 0 & \text{if sensor out of order} \end{cases} \quad (28)$$

3.5 Speech Detection

The goal of speech detection is to measure the auditory interactions of the pilots with the system and detect if they are speaking or not. There is currently no interpretation of the speech signal as such.

The raw data is the amplitude signal $A(t)$ of a microphone, which is integrated in the head set of a pilot (see Fig. 8). Speech detection is based on calculating the power level

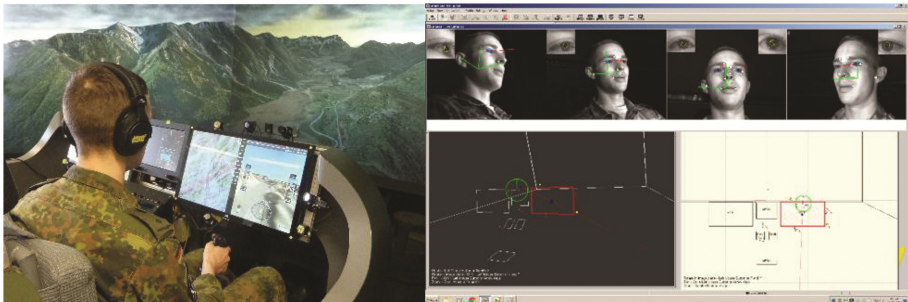


Fig. 8. Integrated gaze tracking system Smart Eye Pro [38].

P by averaging the squared amplitude over a short time interval Δt (here a few hundred milliseconds).

$$P(t) = \frac{1}{\Delta t} \int_{t' = t - \Delta t}^t dt' A^2(t') \quad (29)$$

For detecting speech, a power level P_d is used as the discrimination threshold x_d in Eq. (22):

$$S(P) = \begin{cases} true & \text{if } P \geq P_d \\ false & \text{if } P < P_d \\ unknown & \text{if sensor out of order} \end{cases} \quad (30)$$

Two distinct power levels can be identified. The first one is the signal level if the pilot is speaking P_{signal} , the second one is the noise level P_{noise} if the pilot is not speaking, but the signal is still disturbed by noise from the environment. These parameters result from a calibration, where the power level is recorded for some seconds. The closer the current power level P is to the detection threshold, the worse is the reliability of the measurement:

$$Z(P) = \begin{cases} \frac{P - P_d}{P_{signal} - P_d} & \text{if } P \geq P_d \\ [2mm] \frac{P - P_d}{P_{noise} - P_d} & \text{if } P < P_d \\ [1mm] 0 & \text{if sensor out of order} \end{cases} \quad (31)$$

3.6 Gaze Tracking

For the purpose of getting evidences from visual interactions with the cockpit displays, a commercial gaze tracking system (Smart Eye Pro, Smart Eye AB) has been implemented in the simulator [38]. It is a video-based system, which consists of four cameras for the left pilot and four cameras for the right pilot running at 60 Hz. The cameras are placed around the MFD (see left image of Fig. 8). The gaze tracking method is based on measuring the cornea reflection in the infrared spectrum. There is no intrusion of the pilots in their working domain. A 3D world model of the simulator displays and the outside view has been implemented in the system [38] (see right image of Fig. 8 in the lower left corner).

The raw sub-symbolic values of the eye tracker are pixel coordinates on a simulator display. For creating evidences, semantic information about the objects the pilots are looking at are necessary. These objects can be the airspeed indicator or tactical symbols in the tactical map and are provided by the MFDs [36]. The semantic information on which display, page and object the pilot is looking on is gained by combining the raw pixel coordinates with the layout on the cockpit displays (Fig. 9).

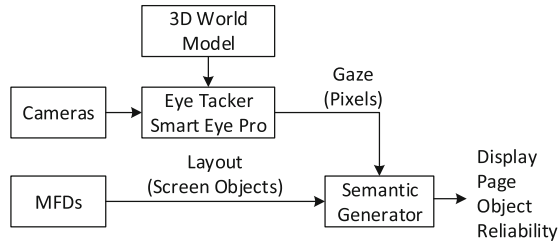


Fig. 9. Gaze tracking system for deriving object oriented semantic data.

In order to take account for measurement inaccuracies, the gaze on the screen is not only described by single pixel coordinates, but rather by a normalized Gaussian distribution over the 2D surface (see Fig. 10):

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-x_0)^2}{\sigma_x^2} + \frac{(y-y_0)^2}{\sigma_y^2} - \frac{2\rho(x-x_0)(y-y_0)}{\sigma_x\sigma_y} \right] \right) \quad (32)$$

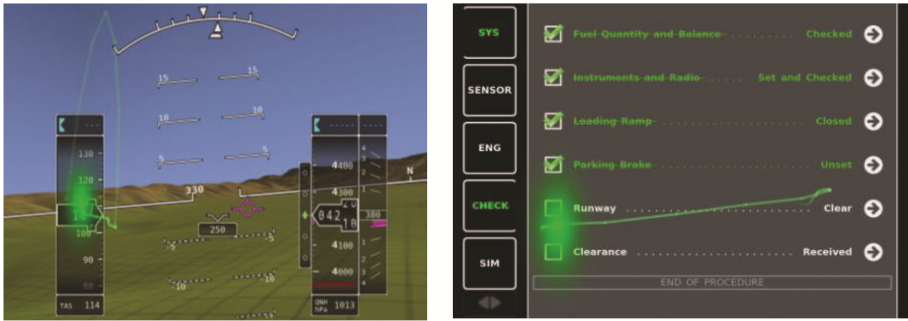


Fig. 10. Gaze tracking on PFD and during processing a check list.

The raw pixel coordinates of the current gaze coordinates delivered by the gaze tracker are given by x_0 and y_0 . The widths of the gaze distribution in the x- and y- screen axis directions are expressed by two different standard deviations σ_x and σ_y . The Pearson correlation coefficient ρ indicates the correlation between the two axes and therefore describes the shear deformation of the gaze spot. These parameters are not constant over the screen, but depend on the current gaze position x_0 and y_0 :

$$\sigma_x = \sigma_x(x_0, y_0) \quad \sigma_y = \sigma_y(x_0, y_0) \quad \rho = \rho(x_0, y_0) \quad (33)$$

We are using bilinear models to describe the dependence of the parameters on the screen position:

$$\begin{aligned} \sigma_x(x_0, y_0) &= \beta_{x1} + \beta_{x2}x_0 + \beta_{x3}y_0 + \beta_{x4}x_0y_0 \\ \sigma_y(x_0, y_0) &= \beta_{y1} + \beta_{y2}x_0 + \beta_{y3}y_0 + \beta_{y4}x_0y_0 \\ \rho(x_0, y_0) &= \beta_{r1} + \beta_{r2}x_0 + \beta_{r3}y_0 + \beta_{r4}x_0y_0 \end{aligned} \quad (34)$$

The constant parameters of these bilinear models are gained during the calibration process of the gaze tracking system individually for each pilot. During this calibration, multiple gaze points on the screen are sampled. Hereby, a least square problem is solved. This is done by a QR-decomposition contained in the Eigen C++ library [39].

Every object displayed on a screen is represented by a polygon of pixel coordinates (see Fig. 11). For determining a symbolic hypothesis S , if the pilot is looking on this object or not, and a corresponding scalar reliability value Z (see Sect. 3.2), the integral of the gaze distribution over each single screen object is calculated.

$$I = \iint_{Screen\ Object} dx\ dy\ f(x, y) \tag{35}$$

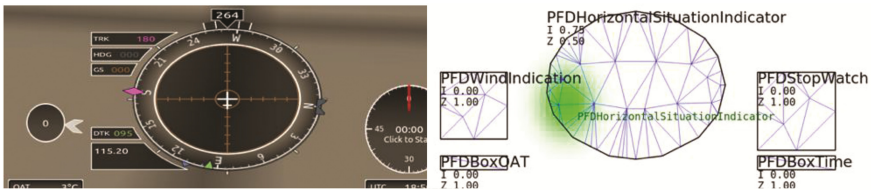


Fig. 11. Example for estimating the reliability of a gaze tracking evidence. The left image shows an excerpt from the PFD and the right image the triangulated screen objects with the Gaussian gaze distribution on the horizontal situation indicator (green spot). (Color figure online)

This integration is calculated technically, by decomposing the polygonal screen objects into triangles with a Delaunay triangulation algorithm [40] [41, pp. 1131–1141], and integrating over the individual triangles by using a Gauss-Legendre quadrature [41, pp. 179–200]. Figure 11 show an example for screen objects on the primary flight display (PFD).

With Eqs. (22) and (23) ($x_{min} = 0, x_{max} = 1, x_d = 0.5, x = I$), the symbolic value and reliability are given by:

$$S(I) = \begin{cases} true & \text{if } I \geq 0.5 \\ false & \text{if } I < 0.5 \end{cases} \tag{36}$$

$$Z(I) = \begin{cases} 2(I - 0.5) & \text{if } I \geq 0.5 \\ -2(I - 0.5) & \text{if } I < 0.5 \end{cases} \tag{37}$$

3.7 Touchscreen Input Detection

The multi-function displays (MFDs) in the simulator cockpit [36] are equipped with a touch-sensitive surface (Fig. 12). The pilot can interact with the MFDs by tapping on the surface with one or more fingers. Furthermore, the displays support pan, swipe and pinch gestures.



Fig. 12. Multi-touch multi-function Display (MFD) in the simulator cockpit.

Touchscreen inputs may be faulty since it is not easy to hit small buttons or symbols on the touchscreen with the finger. Therefore, an error model for touchscreen inputs is assumed. Similar to the error model for the gaze tracker, the touch point is not only a single point, but rather an axial symmetric Gaussian distribution:

$$f(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{r(x, y)^2}{2\sigma^2}\right) \quad (38)$$

$$r(x, y) = \sqrt{(x - x_0)^2 + (y - y_0)^2}$$

The measured coordinates of the touch event are x_0 and y_0 . The parameter σ defines the width of this distribution. This parameter is determined during calibration as the standard deviation of some touch samples. The calculation for deriving the symbolic value S and the reliability Z on the basis of this distribution is the same as for the gaze tracking system (see Sect. 3.6).

3.8 Situation Assessment

Not only the current observations from monitoring the pilots is important, but also the task context given by the situation. Therefore, the tactical situation and system states are also collected to feed the evidential reasoning algorithm. Beside threats from enemy forces, we are considering flight states and flight phases of the helicopter, states of the cockpit-displays, as well as states of the UAVs and their sensors. Weather, time of day and other environmental quantities are neglected in the current setup. Evidences, of the context are mainly used to create rejecting evidences (9). For simplicity, no real situation sensors are modeled at the moment but all observed quantities are taken as perfect reliable observations similar to button presses.

4 Current Status and Future Work

The method presented in Sect. 2 has been implemented in our helicopter mission simulator as described in Sect. 3. The developed software modules are part of a workload-adaptive associate system, which uses a common task model as central representation

of task oriented knowledge [18]. To determine all the possible tasks, which can occur during a mission, a task analysis has been performed together with helicopter pilots and UAV operators of the German Armed Forces [42].

The task model contains currently 226 tasks in total and 85 of them are used for activity determination. The other tasks are abstract and used for hierarchical structuring the model and for other functions of the associate system like mission planning. Currently, there are about 1300 automatically generated observables. 440 of them are linked to the 85 tasks in the task model as evidences by a knowledge engineer. With the help of an inheritance mechanism between tasks in the task model about 730 evidences are generated in total. This results in an average of 8.6 evidences per task. Details about abstract tasks and inheritance relations are described in [18].

The total processing time of the algorithms running in the simulator takes between 40 and 80 ms on a state-of-the-art personal computer and therefore complies with the soft real-time requirement stated in Sect. 2.1.

The system is currently being hardened and the parameters are tuned. In the future, full mission simulations with aviators of the German Armed Forces are planned to evaluate this method for activity determination and the closed loop with activity determination as part of the embracing associate system.

References

1. Uhrmann, J., Strenzke, R., Schulte, A.: Task-based guidance of multiple detached unmanned sensor platforms in military helicopter operations. COGIS (COGNitive systems with Interactive Sensors) (2010)
2. Whittle, R.: MUM-T is the word for AH-64E: Helos Fly, Use Drones. *Breaking Defense* (2015). <http://breakingdefense.com/2015/01/mum-t-is-the-word-for-ah-64e-helos-fly-use-drones/>. Accessed 4 Dec 2016
3. Cooley, M.: Human centred systems: An urgent problem for systems designers. *AI Soc.* **1**(1), 37–46 (1987)
4. Wiener, E.L., Curry, R.E.: Flight-deck automation: promises and problems. *Ergonomics* **23**(10), 995–1011 (1980)
5. Wiener, E.L.: Human Factors of Advanced Technology (“Glass Cockpit”) Transport Aircraft. NASA Contractor Report No. 177528, Moffett Field, CA (1989)
6. Sarter, N.B., Woods, D.D.: How in the world did we ever get into that mode? mode error and awareness in supervisory control. *Human Factors J. Hum. Factors Ergonomics Soc.* **37**(1), 5–19 (1995)
7. 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 Associates, Mahwah (1996)
8. Billings, C.E.: Benefits and costs of aviation automation. *Aviation Automation: The Search for a Human-Centered Approach*, pp. 181–218. Lawrence Erlbaum Associates, Mahwah (1997)
9. Onken, R., Prévot, T.: CASSY - cockpit assistant system for IFR Operation. In: 19th ICAS Congress Proceedings, vol. 3, pp. 2598–2608, Anaheim, CA (1994)
10. Walsdorf, A., Onken, R., Eibl, H., Helmke, H., Suikat, R., Schulte, A.: The crew assistant military aircraft (CAMA). In: *The Human-Electronic Crew: The Right Stuff? 4th Joint GAF/RAF/USAF Workshop on Human-Computer Teamwork*, Kreuth (1997)

11. Miller, C.A., Hannen, M.D.: User acceptance of an intelligent user interface: a rotorcraft pilot's associate example. In: Proceedings of the 4th International Conference on Intelligent User Interfaces, IUI 1999, pp. 109–116, Redondo Beach, CA (1999)
12. Onken, R., Schulte, A.: System-Ergonomic Design of Cognitive Automation: Dual-Mode Cognitive Design of Vehicle Guidance and Control Work Systems. System-Ergonomic Design of Cognitive Automation. Springer, Heidelberg (2010)
13. Maiwald, F., Schulte, A.: Enhancing military helicopter pilot assistant systems through resource adaptive dialogue management. In: Vidulich, M.A., Tsang, P.S., Flach, J.M. (eds.), *Advances in Aviation Psychology*. Ashgate Studies in Human Factors and Flight Operations, pp. 177–196. Ashgate Publishing, Ltd., Farnham (2014)
14. Parasuraman, R., Bahri, T., Deaton, J.E., Morrison, J.G., Barnes, M.: *Theory and Design of Adaptive Automation in Aviation Systems*. NAWCADWAR-92033-60. Naval Air Warfare Center, Aircraft Division, Warminster (1992)
15. Veltman, J.A., Jansen, C.: *The Role of Operator State Assessment in Adaptive Automation*. TNO Defence Security and Safety, Soesterberg (2006). TNO-DV3 20
16. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. *Ergonomics* **58**(1), 1–17 (2015)
17. Schulte, A., Donath, D., Honecker, F.: Human-system interaction analysis for military pilot activity and mental workload determination. In: *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, pp. 1375–1380, Kowloon Tong (2015)
18. Honecker, F., Brand, Y., Schulte, A.: A task-centered approach for workload-adaptive pilot associate systems. In: *Proceedings of the 32rd Conference of the European Association for Aviation Psychology – Thinking High and Low: Cognition and Decision Making in Aviation, Cascais* (2016)
19. O'Donnell, R., Eggemeier, T.: Workload assessment methodology. In: Boff, K., Kaufman, L., Thomas, J. (eds.) *Handbook of Perception and Human Performance*. Wiley, New York (1986)
20. Schmitt, F., Schulte, A.: Mixed-initiative mission planning using planning strategy models in military manned-unmanned teaming missions. In: *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, pp. 1391–1396, Kowloon Tong (2015)
21. Donath, D., Schulte, A.: Behavior based task and high workload determination of pilots guiding multiple UAVs. In: *6th International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences, AHFE 2015*, vol. 3, pp. 990–997. Elsevier, Las Vegas (2015)
22. Butchibabu, A., Sparano-Huiban, C., Sonenberg, L., Shah, J.: Implicit coordination strategies for effective team communication. *Hum. Factors* **58**(4), 595–610 (2016)
23. Eagleman, D.M.: Time and the brain: how subjective time relates to neural time. *J. Neurosci.* **25**(45), 10369–10371 (2005)
24. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38**(2), 325–339 (1967)
25. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, London (1976)
26. Shortliffe, E.H.: *Computer-Based Medical Consultations: MYCIN* (Artificial Intelligence Series). Elsevier Science Ltd., Amsterdam (1976)
27. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc, San Francisco (1988)
28. Hanson, R., Stutz, J., Cheeseman, P.: *Bayesian Classification Theory*. Technical report FIA-90-12-7-01. NASA Ames Research Center. Artificial Intelligence Research Branch, Moffett Field (1991)

29. Lewis, D.D., Ringuette, M.: A comparison of two learning algorithms for text categorization. In: Third Annual Symposium on Document Analysis and Information Retrieval, vol. 33, pp. 81–93 (1994)
30. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. Learning for Text Categorization: Papers from the AAAI Workshop. Technical report WS-98-05, 62, pp. 98–105 (1998)
31. Wittig, T.: Maschinelle Erkennung von Pilotenabsichten und Pilotenfehlern über heuristische Klassifikation. Fortschritt-Berichte VDI. VDI-Verlag, München (1994)
32. Gordon, J., & Shortliffe, E. H. (1984). The Dempster-Shafer theory of evidence. Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, vol. 3, pp. 832–838 (1984)
33. Dempster, A.P.: The Dempster-Shafer calculus for statisticians. *Int. J. Approximate Reasoning* **48**(2), 365–377 (2008)
34. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *Int. J. Approximate Reasoning* **9**(1), 1–35 (1993)
35. Yaghlane, B., Ben, Mellouli, K.: Inference in directed evidential networks based on the transferable belief model. *Int. J. Approximate Reasoning* **48**(2), 399–418 (2008)
36. Brand, Y., Schulte, A.: Human agent interfaces as a key element for the dialog between human crews and cognitive automation. AIAA Infotech@ aerospace, pp. 1–12 (2015)
37. Weinmann, A., Brand, Y., Honecker, F., Meyer, C., Rudnick, G., Ruf, C., Schmitt, F.: Flight and user manual IFS helicopter mission simulator. Institute of Flight Systems. Aerospace Engineering Department. Bundeswehr University Munich, Munich (2016)
38. Mehler, J.: Integration und Evaluierung eines Blickbewegungsmesssystems für Pilot und Kommandant in einen Hubschraubersimulator. Universität der Bundeswehr München, Bachelorarbeit (2014)
39. Jacob, B., Guennebaud, G., et al.: Eigen: C++ template library for linear algebra. Version 3. <https://eigen.tuxfamily.org>. Accessed 27 Jan 2017
40. Delaunay, B.: Sur la sphère vide. A la mémoire de Georges Voronoi. *Bulletin de l'Académie des Sciences de l'URSS, Classe des sciences mathématiques et na* **6**, 793–800 (1934)
41. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes - The Art of Scientific Computing, 3rd edn. Cambridge University Press, New York (2007)
42. Winkler, B.: Design und Evaluierung von militärischen MUM-T-Missionen mittels Expertenwissen als Grundlage für zukünftige Messkampagnen. Universität der Bundeswehr München, Bachelorarbeit (2016)

Assessing Human-Computer Interaction of Operating Remotely Piloted Aircraft Systems (RPAS) in Attitude (ATTI) Mode

Pete McCarthy¹(✉) and Guan Kiat Teo²

¹ Cranfield University, Cranfield, UK

Pete.mccarthy@cranfield.ac.uk

² Transport Safety Investigation Bureau, Singapore

Steven_TEO@mot.gov.sg

Abstract. The addition of relatively cheap, yet accurate and reliable automated flight controllers, to even the most basic sub 20 kg (KG) RPAS/drone, has revolutionised the use of these systems, and made them widely accessible to the general public. Because of this, drone use covering a wide variety of applications has increased in recent years, and is set to continue to increase at an exponential pace. While drone automation allows novices to easily control and operate their aircraft, it can also however create a false sense of confidence, that the drones can be operated with little or even no training at all! When automation fails, however, drone pilots may find themselves having to control their unmanned/remotely piloted aircraft with greatly reduced technological assistance. This mode of operation is known as Attitude (ATTI) Mode and occurs when the flight control system loses Global Positioning System (GPS) accuracy. Currently, in the UK, drone pilots wishing to operate a platform below 20 kg in weight, need to undergo a practical assessment, which requires the drone to be flown in ATTI Mode. However, there is no clear guidance on what test flight profiles they may be asked to be fly. This creates a situation where drone pilots may be subjected to an extremely wide variance of practical assessments. This research consolidates from UK CAA-approved drone operators the types of flight profiles that they had been asked to demonstrate in ATTI Mode during their practical assessments. From all the profiles reported, the seven most frequently reported flight profiles were further analysed to rank their effectiveness in assessing drone pilots' flight operation competency. It has been found that some of these flight profiles are not statistically significantly different from one another. Accordingly, it is proposed that assessors may consider selecting flight profiles that are significantly different to be performed during the practical assessment for a drone pilot, so that time and effort will not be wasted, but more importantly, the assessment of the SUA pilots' competency may become more comprehensive.

Keywords: RPAS · Drone · Human factors · Interaction · Automation

1 Introduction

Unmanned aircraft systems can range from the simplest form of a single unmanned platform programmed to fly a pre-set flight path for a pre-defined duration, to a system.

Comprising of the unmanned aircraft with a full suite of complementing ground equipment that provides real-time command and control of the unmanned platform itself and the payloads it's carrying. Many different terms have been used to try and define different elements of unmanned aircraft systems. At times, a definition may have different meanings when referred to in a different context or by different individuals.

In this study, an unmanned aircraft, other than a balloon or a kite, having a mass of not more than 20 kg without its fuel but including any articles or equipment installed in or attached to the aircraft at the commencement of its flight, is defined as a RPAS or drone.

1.1 RPAS Automation and Typical Operating Modes

The fastest growth of unmanned aircraft for non- military applications is in the class of small drone (sub 20 kg), due to its relatively low cost, yet possessing the capability and technology to meet its intended objectives. Within this class of drone, multi-rotor platforms which are capable of taking off and landing vertically has generated the greatest interest among the community owing to their inherent ability to be launched and recovered from confined locations and ability to perch and stare for extended periods of time (Prior 2013). Some of the most common non-military applications include aerial photography and videography, borders surveillance and simply recreational flying.

Although relatively cheap, drones are equipped with technologies that make flying them fairly easy. For example, they can be programmed to fly along a predefined flight path automatically via a ground- based computer software that is usually very simple and intuitive to operate. Once the flight path is programmed and uploaded to the drone, the automation can make use of the Global Positioning System (GPS) signal to navigate along the planned route accurately.

Besides flying the drone along a predefined flight path based on pre-uploaded data, pilots may also choose to navigate by means of a remote controller. A typical remote controller, a description of the control input and the corresponding motion of the SUA relative to an imaginary person sitting on board the SUA facing forward are shown in Fig. 1.

When the GPS signal received by the drone is sufficiently strong, the aircraft will be able to operate in GPS Mode. This results in the drone flying accurately and smoothly, by comparing the actual position of the drone with the input from the remote controller. For example, when a pilot moves only the pitch control up in GPS Mode, the drone will move forward and any lateral deviation within its operating limits will be compensated for by the automation. This results in the aircraft moving only straight ahead relative to the drone with its vertical distance from the ground being constant. In addition, operating in GPS Mode allows the position (and vertical distance) of the drone to be locked when the inputs of the flight controls (i.e. pitch, roll and yaw) are neutralised.

While automation allows novices to easily control and operate drones in GPS Mode, sometimes without much prior training, it can create a false sense of confidence that the drones can be operated without much training, if any is considered necessary at all! Currently, in the United Kingdom (UK), it is not a requirement for drones below



Fig. 1. A typical drone remote controller

20 kg or its software to be certified airworthy by the European Aviation Safety Agency (EASA) or the Civil Aviation Authority (CAA 2015b). Thus, the reliability of the automation may not be sufficiently high to ensure that the drone will always be able to operate in GPS Mode, or that there will be any fail safe when the GPS signal is not received.

Moreover, certain drone operations may be required to be conducted within built-up areas, or even indoors, and this may have an adverse effect on the reception and integrity of the GPS signal. The availability of the GPS signal may also be subjected to weather such as cloud cover and precipitation. In these abovementioned examples, the drone may still be operated in a lower level of automation known as the Attitude (ATTI) Mode.

In ATTI Mode, the drone is still flown based on input via the remote controller in the same manner as compared to GPS Mode. While the vertical distance of the drone can still be maintained by the onboard automation, the position of the drone will not be automatically locked. In contrast to the example cited earlier when the aircraft was flown in GPS Mode, when a drone pilot moves only the pitch control up in ATTI Mode, the aircraft may veer off its track laterally as it moves forward in response to the control input. This poses a much greater challenge to the drone pilots as their control input will need to be very accurate, precise and very dynamic as the aircraft reacts and responds to environmental conditions such as gusts. In order for the aircraft to maintain the desired track (i.e. straight ahead without any lateral deviation), the pilot will also need to apply a suitable roll input to compensate for the lateral drift. Similarly, when all the flight controls are neutralised, the aircraft may not be able to maintain its hovering position in ATTI Mode. Instead, it will drift from its intended position according to the wind conditions. In order for the drone to hover in a fixed position in ATTI Mode, the pilot will have to continuously apply input to the remote controller as the aircraft is swayed by the environmental elements.

1.2 Differences Between Manned and Unmanned Pilots' Training Requirements

With the removal of pilots from the flying machine, different hazards, which in some ways are greater than those of manned aircraft, are introduced. These novel hazards are not addressed by traditional training regimes of the manned pilots (McCarley and Wickens 2005; Hayhursy et al. 2006). Seated in the cockpit behind the flight controls with an array of panels and displays, a manned aircraft pilot is intimately aware of the surroundings, as well as the state of the aircraft. Cues indicating the aircraft performance and possible failures such as visual and aural alerts, vibrations and smells, are readily available to the manned pilots without the need for a transmission media. On the other hand, when operating RPAS, information pertaining to the performance, orientation, motion and system states of the aircraft become very limited to the drone pilots as they need to be sensed by the on-board automation before being sent through the data transmission medium. In addition, drone pilots are stripped of all vestibular and proprioceptive stimuli, essentially rendering them to operate in “sensory isolation” (Van Erp and Van Breda 1999; McCarley and Wickens 2004; Dalamagjidis et al. 2012; International Society of Air Safety Investigators 2015).

Various technologies have been harnessed to enhance and improve the situational awareness of the drone pilot and the controllability of the aircraft. For example, being physically separated from the platform, it is very difficult for drone pilots to detect that the aircraft is encountering turbulence. However, the turbulence can be detected by the aircraft automation and sent to the ground control station or remote controller via the wireless transmission medium. The information can then be conveyed to the pilot through, for example, the vibrating of the remote controller to create awareness of the turbulence (Calhoun et al. 2002). Auditory alerts and visual indicators that are normally available to pilots of manned aircraft can also be presented to the drone pilots as a method of alerting operators to system failures enabling better human performance as compared to using only visual indication to reflect systems status (Dixon et al. 2003; Wickens 2010).

Although automation such as automatic navigation, remote controlling in GPS Mode and the transmitting of sensory information to the drone pilots has enabled RPAS to be operated relatively easily, it cannot be depended upon solely to ensure that safe and reliable operation is always maintained due to the possibilities of malfunctioning automation and degraded GPS signal. The availability, accuracy, and timeliness of this sensory and system information is heavily dependent on the automation software which currently are not being demonstrated to or certified by any aviation authorities to be sufficiently reliable.

In addition to the on-board automation and sensors' limitations in their reliabilities, the quality and timeliness of the data presented to the drone pilot on the ground will also be constrained by the bandwidth and quality of the communications link between the drone and ground control station or remote controller. (McCarley and Wickens 2004). Data link bandwidth limits, for example, will limit the temporal resolution spatial resolution and field of view of the visual displays presented to drone pilots on the ground and may adversely affect the judgment and decision-making process of the pilot (Van Erp 1999). Besides the quality of the data, there is always a

time delay from the transmission of the data to the time they are received by the drone pilot on the ground. This further compounds the difficulty of the drone pilot in receiving up to date and accurate information and status of the aircraft in order to maintain control (Gawron 1998). Other than causing drone pilots to always receive slightly outdated information, data transmission delays reduce the time available for the pilots to process the information and respond with the most appropriate control input as fast as possible.

Rogers et al. (2004) and Tvaryanas et al. (2006) found that as high as 68% of the accidents and incidents involving unmanned aircraft can be linked to the lack of situation awareness. Although the drone automation, GPS signal and the quality of the communication link are usually very reliable, pilots should refrain from over-relying on full functioning automation to operate their aircraft. The ease of operating a drone in GPS Mode and the seemingly reliable (at least most of the times) automation can easily lead pilots to excessively trust and over rely on the automation. This over trust in automation can in turn slowly and eventually erode their skills required to manually operate the aircraft (Parasuraman and Riley 1997). If the drone pilot's skills do get eroded and when the situation arises that require the pilot to quickly take control of the aircraft without full automation (i.e. in ATTI Mode), they may not be able to assess the situation quickly enough and take the most appropriate recovery actions competently or confidently. Instead, the pilots in these situations may find themselves being left "out of the loop" and not be able to regain safe control of the aircraft (Billings 1991; Wickens and Hollands 2000; Mouloua et al. 2001; Sharma and Chakravarti 2005).

Regardless of the status of on-board automation, integrity of GPS signal received and quality of data transmission link, drone pilots remain responsible for the safe operation of their aircraft. So, it is important that pilots are able to competently regain safe operation of the aircraft via the lowest level of control, i.e. the remote controller and operating in Attitude (ATTI) Mode, when necessary.

In a study conducted by the United States Department of Defence, the accident rate of Unmanned Aerial Vehicles (UAV) can be as much as 100 times higher compared to that of manned aircraft (Department of Defence 2001, Schaefer 2003). A significant percentage of the UAV accidents has been attributed to human errors which in several cases can be attributed to inexperience (Williams 2004; Damalagjidis et al. 2012) and operating the unmanned aircraft via remote controllers (Williams 2004; McCarley and Wickens 2005; Williams 2006).

Human factors dissimilar from those normally experienced by pilots operating manned aircraft, including but not limited to sensory deprivation and motion (or the lack of it) inconsistent with the attitude of the aircraft being controlled, place unique physical and mental demands on the drone pilot (McCarley and Wickens 2005; ICAO 2011). One of the key difficulties that drone pilots may face when operating the aircraft is that there is an inconsistent mapping between the movement of the remote controller and the relative response of the aircraft, especially when a part of the aircraft other than its tail is facing the pilot. For example, when the aircraft nose is facing the pilot handling the remote controller, an input to roll the aircraft to the left will cause the aircraft to appear to roll to the right from the perspective of the pilot. This inconsistent mapping of the drone's movement with respect to the pilot is a violation of the human

factors principle of motion compatibility and may place high cognitive demands on the drone pilot (Wickens and Holland 2000; McCarley and Wickens 2005).

As drones transit into ATTI Mode, either due to the degradation of automation, loss of GPS signal or intentionally in order to fit certain types of operation, the difficulty in controlling the aircraft accurately and precisely increases tremendously as the position of the aircraft will be subjected to deviation caused by environmental factors such as wind. However, drone pilots are still expected to maintain safe operation by maintaining visual contact with the aircraft and safely manoeuvre the drone via the remote controller (Stevenson et al. 2015).

It may be advantageous for a set of flight profiles to be identified as relevant and important for drone pilots to demonstrate satisfactorily during their practical assessments in order for them to be recommended by flight examiners for the granting of the PFAW by the CAA. This way, a more standardised scope of the practical assessment of drone pilots with minimum variance between assessments conducted by different flight examiners can be established.

2 Methodology

2.1 Identifying Current Practice for Attitude (ATTI) Mode Assessment

In order to solicit information on how the assessment of drone pilots operating in Attitude (ATTI) Mode is currently being performed in the United Kingdom (UK), a survey form was sent to the entire population of 20 National Qualified Entities (NQE – flight examiners) authorised by the Civil Aviation Authority (CAA) CAA 2015a) in an attempt to find out what are the typical flight profiles they require drone pilots to perform during the practical assessments they conduct. However, only two of the 20 NQEs responded to the survey. Since the sample size cannot be representative of the NQE population, the two responses were not used and this research looked to the commercial drone operators in the UK as the source of information.

The UK CAA publishes a list comprising of drone operators that have been approved to perform commercial or official drone operations in the UK. All of the 1557 drone operators listed (at the time of this study) would have been subjected to at least one practical assessment by a NQE in partial fulfillment of the requirements to be granted a Permission for Aerial Work (PFAW). A similar survey was therefore sent to the entire population of CAA-approved drone operators in the UK. The drone operators were asked to recall and report the flight profiles that they had been asked to demonstrate in ATTI Mode during their practical flight assessments with the NQEs. There is no limit to the number of flight profiles each drone operator may report. Every reported flight profile is recorded as one count. Similar flight profiles, however, are grouped together and the total number of reports would be recorded. For example, if one operator reported a flight profile as flying in a circular path around an object and another operator reported a flight profile as flying a square path around a tree, these two profiles would be grouped together as a single flight profile as ‘flying around an object’ and two counts would be recorded.

The flight profiles consolidated based on the survey responses were then sorted according to the number of times they were mentioned in the survey.

A total of 31 responses were received from the survey sent to CAA-approved drone operators. From the responses, 18 different flight profiles were reported to have been asked to be performed in ATTI Mode by drone pilots during their practical flight assessments. The list of all the flight profiles reported are shown in Table 1 and sorted in decreasing number of times they were reported. The top five most common flight profiles identified from the survey were used as a basis to conduct the second survey.

Table 1. Flight profiles in ATTI Mode drone pilots demonstrated during practical assessments conducted by NQEs

Flight profiles in ATTI Mode	Number of reports
General control	15
Fly around a fixed object (with camera always pointing to object	12
Figure-of-8	11
Landing	9
Flying with the drone in an orientation other than its tail facing the pilot	7
Controlled hover	6
Take-off	6
Emergency departures	3
45° ascent or descent	3
Sudden gusting winds	2
Level circuit	3
Recovering from a Fail Safe or Return Home command	2
360° turn	2
Low level fly-by	1
Rising fly-by	1
High altitude loss of GPS	1

As listed in Table 1, the most common flight profile reported by drone operators was ‘general handling’. As this is a very generic and vague profile to determine its effectiveness of, three flight profiles that had been grouped together under the ‘general handling’ profile were used to solicit input from the participants of the second survey where participants will be asked to rate the effectiveness of various flight profiles. These three profiles are ‘following a line’, ‘following a route’ and ‘recovering from wind’. In summary, the flight profiles that are included in the second survey are as follow:

- Following a line,
- Following a route,
- Figure-of-8,
- Recovering from wind,

- Circling an object,
- Drone orientated other than tail facing the pilot (hereafter referred to as ‘non-tail-facing pilot’)
- Landing.

2.2 Rating Effectiveness of Identified ATTI Mode Flight Profiles

A second survey was conducted where participants were asked to rate the effectiveness of the seven flight profiles identified from the first survey. For each flight profile, a short video clip of a SUA being flown in accordance to the profile was produced. In each video, an inset was included to illustrate to the survey participants the corresponding input and coordination required from the drone pilot.

A snapshot of one of the video clips is shown in Fig. 2.

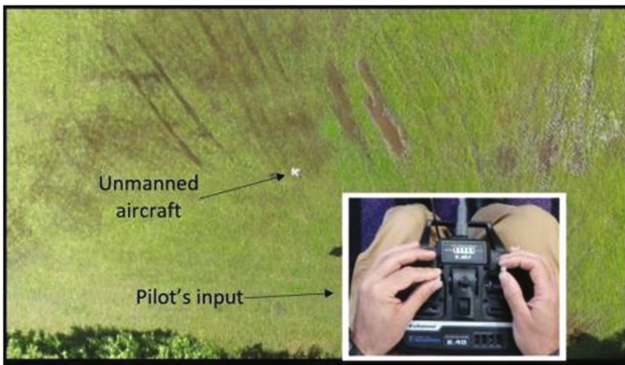


Fig. 2. Snapshot of a video clip from the second survey showing a drone flying various flight profiles with inset illustrating the corresponding pilot input and coordination's

At the beginning of the second survey, the participants were provided with an explanation of the various inputs of the remote controller used in the survey and an explanation of what ATTI Mode means and entails. After watching each video, participants were then asked to rate the effectiveness of the flight profile when used to assess the competence of drone pilots.

It was decided not to send this survey to the targeted populations in the first survey, i.e. NQEs and CAA-approved drone operators in the UK, so as to avoid any prejudiced or biased responses. For example, a drone pilot who is unable to proficiently perform a particular flight profile included in the second survey may deliberately rate that corresponding profile as being not effective at all, regardless of its actual effectiveness. Instead, post-graduate students in the UK were invited to participate in this survey. This sample group is considered to be neutral and unbiased, thus providing a more objective evaluation of the survey questions.

2.3 Statistical Analysis of Survey Data

A total of 28 fully completed responses were received from the second survey. These responses were first tested for normality using the Kolmogorov-Smirnov test. It was found that the participants' responses for all the flight profiles were normally distributed. From the histograms, it was also observed that the distributions of the effectiveness ratings for the 'circling an object' and 'landing' flight profiles were negatively skewed.

The mean effectiveness ratings of the flight profiles, as shown in Table 2 in descending order of mean effectiveness, indicate that the 'landing' profile (mean = 4.14, SD = 1.01) was rated the most effective profile whereas the 'following a line' profile (mean = 2.68, SD = 0.86) was considered the least effective.

Table 2. Mean and standard deviation (SD) of effectiveness of flight profiles

Flight profile	Mean	SD
Landing	4.14	1.01
Circling an object	4.04	0.92
Non-tail-facing pilot	3.79	0.88
Figure-of-8	3.50	0.92
Following a route	3.00	1.02
Recovering from wind	2.93	1.02
Following a line	2.68	0.86

The data was then subjected to the Mauchly's Test of Sphericity, which indicated that the assumption of sphericity had not been violated ($\chi^2(20) = 24.689$, $p = 0.217$). Thus, no correction for the degree of freedom was required.

The mean effectiveness ratings of the seven flight profiles were then tested to investigate if their differences are significant by an ANOVA with repeated measures test with sphericity assumed. The test results revealed the means were statistically significantly different ($F(6, 162) = 13.17$, $p < 0.005$).

A post-hoc pairwise comparison of the means of the effectiveness ratings using the Bonferroni correction was further conducted. Results of the comparison revealed that the significant differences exist only between the flight profiles listed below. Details of the pairwise comparison results are shown in Table 3.

- Following a line and Figure-of-8 ($p < 0.05$)
- Following a line and Circling an object ($p < 0.05$)
- Following a line and Non-tail-facing pilot ($p < 0.05$)
- Following a line and Landing ($p < 0.05$)
- Following a route and Circling an object ($p < 0.05$)
- Following a route and Non-tail-facing pilot ($p < 0.05$)
- Following a route and Landing ($p < 0.05$)
- Figure-of-8 and Landing ($p < 0.05$)
- Recovering from wind and Circling an object ($p < 0.05$)
- Recovering from wind and Non-tail-facing pilot ($p < 0.05$)
- Recovering from wind and Landing ($p < 0.05$).

Table 3. Post-hoc pairwise comparison test results on effectiveness ratings of flight profiles in ATTI Mode

Flight profiles		Mean difference	Std error	Significance
Following a line	Following a route	0.321	0.200	1.00
Following a line	Figure-of-8	0.821	0.225	0.023*
Following a line	Recovering from wind	0.250	0.210	1.000
Following a line	Circling an object	1.357	0.213	<0.001*
Following a line	Non-tail-facing pilot	1.107	0.201	<0.001*
Following a line	Landing	1.464	0.227	<0.001*
Following a route	Figure-of-8	0.500	0.196	0.350
Following a route	Recovering from wind	0.071	0.241	1.000
Following a route	Circling an object	1.036	0.249	0.006*
Following a route	Non-tail-facing pilot	0.786	0.181	0.004*
Following a route	Landing	1.143	0.216	<0.001*
Figure-of-8	Recovering from wind	0.571	0.264	0.834
Figure-of-8	Circling an object	0.536	0.221	0.47
Figure-of-8	Non-tail-facing pilot	0.286	0.240	1.000
Figure-of-8	Landing	0.643	0.172	0.019*
Recovering from wind	Circling an object	1.107	0.274	0.008*
Recovering from wind	Non-tail-facing pilot	0.857	0.234	0.023*
Recovering from wind	Landing	1.214	0.288	0.005*
Circling an object	Non-tail-facing pilot	0.250	0.197	1.000
Circling an object	Landing	0.107	0.208	1.000
Non-tail-facingpilot	Landing	0.357	0.237	1.000

*Statistically significantly different.

3 Discussion

The test for normality of the data consolidated from the second survey showed that the ratings of the effectiveness of all the flight profiles surveyed were normally distributed. In addition, visual examination of the histograms revealed that the distributions for the ‘circling an object’ and ‘landing’ flight profiles were negatively skewed. One possible explanation for the negative skew of the flight profile ‘circling an object’ is that not only does the drone pilot need to maintain the aircraft within visual range while accurately controlling its flight path just like any other flight profiles, the object that is being flown around needs to remain constantly within the field of view of the optical camera attached to the drone. This means that the drone pilot will have to allocate additional cognitive resources to another critical task of maintaining the target within the camera’s view at all times. Since this flight profile requires additional cognitive resources attending to an additional task (i.e. monitoring the video imagery sent from the drone camera) when compared to the other flight profiles, the successful execution of the ‘circling an object’ flight profile may, on the average, be considered as a very highly effective flight profile to assess the competency of a drone pilot. This may have

led more participants to rate this profile as highly effective, resulting in the negative skew of the histogram.

The other flight profile with a negatively skewed histogram was the 'landing' profile. Incidentally, the 'landing' flight profile also scored the highest mean effectiveness rating (mean = 4.14, SD = 1.01). In the event of a drone transiting to ATTI Mode due to, for example, a malfunction of automation, a drone pilot may choose to terminate the flight as soon as practicable, thus not be required to perform any further complicated flight profiles such as following a predetermined route. However, the aircraft almost always has to be landed safely. The fact that drone pilots are expected to always land their aircraft safely in ATTI Mode may have influenced most of the survey participants to rate the 'landing' flight profile as the most effective flight profile among those that are included in the second survey, resulting in the 'landing' flight profile having a negatively skewed histogram and also scoring the highest mean effectiveness.

The 'following a line' flight profile has been rated as the least effective flight profile (mean = 2.68, SD = 0.86) to assess the competency of drone pilots. One possible explanation for this is that this profile may have seemed to be very easy to the survey participants since they are not able to fully experience the challenges involved to ensure that the drone maintains the planned line. These challenges may include environmental conditions such as the sun glare and wind. Also, the survey participants may also have opined that 'following a line' profile was similar to, and may be thus considered as a subset of, the profile 'following a route'. As a result, the 'following a line' profile is rated as the least effective, and possibly even be considered irrelevant if the 'following a route' profile is being considered, such as it is in this survey. The effectiveness of some of the flight profiles was statistically found to be not significantly different. For example, the flight profiles 'figure-of-8' and 'non-tail-facing pilot' are not significantly different from each other. However, the controls and coordination required to execute both these flight profiles can be perceived to be rather different. This may lead NQEs to require a drone pilot to demonstrate these two flight profiles in a single practical assessment session. The effect of this, at the very least, may be a waste of time and effort for drone pilots to perform two flight profiles that are not statistically significantly different in demonstrating their competency. At the other end of the spectrum, if all the flight profiles that NQEs require drone pilots to execute are not significantly different, for example only the 'figure-of-8' and 'non-tail-facing pilot' flight profiles were asked to be demonstrated during the practical assessments, the evaluation of the latter's competence may not be sufficiently comprehensive.

With the knowledge of the mean effectiveness and significant difference, or the lack of it, between flight profiles, NQEs can better design a practical assessment that is more effective by employing profiles that are highly effective and significantly different. For example, the 'landing' profile may be always included in the practical assessment since it has been rated as the most effective profile. It may then not be necessary for the second and third most effective flight profiles, i.e. 'circling an object' and 'non-tail-facing pilot', to be performed during the assessment since these two profiles had been found to be not statistically significantly different from the 'landing' profile.

While all the remaining flight profiles were found to be statistically different from the 'landing' profile, there is no significant difference between 'figure-of-8' and 'following a route', between 'figure-of-8' and 'recovering from wind', nor between 'figure-of-8' and 'circling an object'. Thus, the 'figure-of-8' profile may be considered to be included in the practical assessment of drone pilots instead of 'following a route' and 'recovering from wind', since it has been rated as the most effective among these three flight profiles.

Although the 'following a line' profile has been rated as the least effective flight profile, it was found to be statistically significantly different from both the 'landing' and 'figure-of-8' profiles. Thus, NQEs may also consider including the 'following a line' flight profile in the practical assessment of drone pilots in order to assess a wider range of skills and competency.

Accordingly, it is considered that a practical assessment of drone pilots requiring candidates to demonstrate the flight profiles 'landing', 'figure-of-8' and 'following a line' in ATTI Mode can most comprehensively assess drone pilots in the most effective and efficient manner.

4 Conclusion

Currently, National Qualified Entities (NQEs) are requiring drone pilots to demonstrate a very wide array of flight profiles when assessing their competency as a prerequisite to the latter being granted a Permission for Aerial Work (PFAW) by the Civil Aviation Authority (CAA). This large variance of practical assessments is due to the lack of guidance on the recommended flight profiles that NQEs may stipulate drone pilots to perform during their practical assessments. This may result in a waste of time and effort, as there is a possibility that only flight profiles that are not significantly different would be asked to be performed. In such cases, the assessment of the drone pilot's competency may also be not sufficiently comprehensive to assess a wider range of skills and competency.

CAA-approved drone operators in the UK have been asked to describe the flight profiles they had been asked to demonstrate during their practical assessments through a survey. The effectiveness of the reported flight profiles to assess SUA pilots' competency in operating their unmanned aircraft were then rated by post-graduate students in the UK through a second survey. Statistical analysis of the responses from the second survey reflected the mean effectiveness of each flight profiles to assess the competency of drone pilots and also revealed that not all the flight profiles are significantly different from one another.

With this information, NQEs may consider selecting flight profiles that are highly effective and significantly different from one another to be performed by SUA pilots during practical assessments. A possible set of flight profiles that fit these conditions comprises of the flight profiles 'landing', 'figure-of-8' and 'following a line'. By designing practical assessments of SUA pilots based on this principle, time and effort would not be spent on flight profiles that are basically testing similar skill sets. More importantly, the assessment of the SUA pilots' competency may be more comprehensive.

4.1 Further Research

In order to identify a set of flight profiles that drone pilots are currently required to perform in Attitude (ATTI) Mode during their practical assessments, a survey was first distributed to ask SUA operators in the United Kingdom (UK) what were the flight profiles they had to perform during their practical assessment with a National Qualified Entity (NQE). Out of the 18 types of flight profiles consolidated, the top five (which were eventually expanded to seven) were used in a second survey to solicit the effectiveness of each profile in assessing a drone pilot's competency. In doing so, the opportunity is lost in evaluating the remaining 13 flight profiles which, although currently used less often by NQEs, could well be more effective than the ones analysed in this research. There is also the possibility that a new flight profile could be designed to be more effective in assessing drone pilots' competency than all those that had been identified. The scope of this research could possibly be expanded to analyse all 18 flight profiles that had been reported to be required by NQEs to be performed during the practical assessment. Also, new flight profiles may be designed and included in the evaluation.

The NQEs and CAA-approved drone operators in the UK were not asked to participate in the second survey in order to avoid prejudice and bias in the collected data. Instead, the sample group of the second survey consisted of post-graduate students in the UK. While this sample group is able to provide an independent and fair response, their relative inexperience in drone flight operations may affect their perception of the input, coordination, and difficulty level of the flight profiles presented to them in the survey through the video clips. Similar future research could be performed with sample groups that include professionals having relevant drone expertise who will still be able to appraise the survey questions without prejudice or bias. An example of such a sample group may be representatives from the CAA.

In addition to asking the survey participants to rate the effectiveness of the flight profiles presented to them, future surveys may also include a free text field for participants to provide the reasons for their ratings and their comments, if any. With this additional information, the explanations of the histogram distribution and analysis of the data may be further supported and substantiated. For example, although it has already been shown statistically that the effectiveness of 'following a line' and 'following a route' is not significantly different and a possible explanation was discussed, comments from the survey participants may further reinforce the statistics results and analysis.

Much effort was put into this research to help the survey participants understand the challenges involved in operating the drone in ATTI Mode using the remote controller. In order to allow survey participants to fully appreciate the difficulties and skills involved, it may be considered in future research for participants to operate a drone, either a real one or using a simulator, instead of watching a video clip.

References

- Billings, C.E.: Human-centered aircraft automation: a concept and guidelines. National Aeronautics and Space Administration, Ames Research Center, Moffett Field (1991)
- Calhoun, G.L., Draper, M.H., Fontejon, J.V.: Utility of a tactile display for cueing faults. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 2144–2148. SAGE Publications (2002)
- Civil Aviation Authority. Apply for a permission to fly drones for commercial work. UK Civil Aviation Authority (2015). <https://www.caa.co.uk/Commercial-indus-try/Aircraft/Unmanned-aircraft/Apply-for-a-permission-to-fly-drones-for-commercial-work/>
- Dalamagjidis, K., Valavanis, K.P., Piegl, L.A.: On Integrating Unmanned Aircraft Systems into the National Air-Space System, 2nd edn. Springer, Dordrecht (2012)
- Department of Defense. Unmanned aerial vehicle roadmap. General Printing Office, Washington (2001)
- Dixon, S.R., Wickens, C.D., Chang, D.: Comparing quantitative model predictions to experimental data in multiple-UAV flight control. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 104–108. SAGE Publications (2003)
- Gawron, V.J.: Human factors issues in the development, evaluation and operation of uninhabited aerial vehicles. In: Proceedings of the Association for Unmanned Vehicle Systems International, AUVSI 1998, pp. 431–438. AUSVI, Alabama (1998)
- Hayhurst, K.J., Maddalon, J.M., Miner, P.S., DeWalt, M.P., McCormick, G.F.: Unmanned aircraft hazards and their implications for regulations. In: 25th Digital Avionics Systems Conference, pp. 5B1-1–5B1-12 (2006)
- International Civil Aviation Organisation: Unmanned Aircraft Systems: Circular 328. International Civil Aviation Organisation, Montreal (2011)
- International Society of Air Safety Investigators. Unmanned aircraft system handbook and accident/incident investigation guidelines. International Society of Air Safety Investigators, Virginia (2015)
- McCarley, J.S., Wickens, C.D.: Human factors implications of UAVs in the national airspace. University of Illinois, Illinois, Aviation human factors division (2005)
- McCarley, J.S., Wickens, C.D.: Human factors concerns in UAV flight. University of Illinois at Urbana - Champaign Institute of Aviation, Aviation Human Factors Division (2004)
- Mouloua, M., Gilson, R., Daskarolis-Kring, E., Kring, J., Hancock, P.: Ergonomics of UAV/UCAV mission success: considerations for datalink, control, and display issues. In: Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting, pp. 144–148. Human Factors and Ergonomics Society, Minnesota (2001)
- Parasuraman, R., Riley, V.: Humans and automation: use misuse disuse abuse. *Hum. Factors J. Hum. Factors Ergonomics Soc.* 39(2), 230–253 (1997)
- Prior, S.D.: Small remotely piloted aircraft systems. *Defence global*, pp. 65–66 (2013)
- Rogers, B.M., Palmer, B., Chitwood, J.M., et al.: Human-systems issues in UAV design and operation. Wright Patterson AFB, Ohio (2004)
- Schaefer, R.: Unmanned aerial vehicle reliability study. Office of the Secretary of Defense, Washington (2003)
- Sharma, S., Chakravarti, D.: UAV operations: An analysis of incidents and accidents with human factors and crew response management perspective. *Indian J. Aerosp. Med.* 49(1), 29–36 (2005)

- Stevenson, J.D., O'Young, S., Rolland, L.: Assessment of alternative manual control methods for small unmanned aerial vehicles. In: 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015, pp. 952–959. Applied Human Factors and Ergonomics (2015)
- Tvaryanas, A.P., Thompson, W.T., Constable, S.H.: Human factors in remotely piloted aircraft operations: HFACS analysis of 221 mishaps over 10 years. *Aviat. Space Environ. Med.* **77**(7), 724–732 (2006)
- Van Erp, J.B.F., Van Breda, L.: Human factors issues and advanced interface design in maritime unmanned aerial vehicles: a project overview. TNO Human Factors Research Institute, Soesterberg (1999)
- Wickens, C.D.: Multiple resources and performance prediction. *Theor. Issues Ergonomics Sci.* **3**(2), 159 (2010)
- Wickens, C.D., Hollands, J.G.: *Engineering Psychology and Human Performance*. Prentice Hall, Atlantic City (2000)
- Williams, K.W.: A summary of unmanned accident/incident data: human factor simplifications. Federal Aviation Administration, Washington (2004)
- Williams, K.W.: Human factors implications of unmanned aircraft accidents: flight control problems. Federal Aviation Administration, Washington (2006)

Multi-UAV Based Helicopter Landing Zone Reconnaissance

Information Level Fusion and Decision Support

Marc Schmitt^(✉) and Peter Stütz

Institute of Flight Systems (IFS),
University of the Bundeswehr Munich (UBM), Neubiberg, Germany
{marc.schmitt,peter.stuetz}@unibw.de

Abstract. This article presents an information fusion and decision-support system for the multi-UAV based landing zone reconnaissance and landing point evaluation in manned-unmanned teaming (MUM-T) helicopter missions. For this, numerous and heterogeneous data from variety of sensors must be gathered, fused and evaluated. However, payload capacity and on-board processing capabilities are often restricted. Thus, the teaming of multiple unmanned aerial vehicles (UAVs) offers a promising way to overcome these limitations and allows to benefit from heterogenous sensor payloads. Furthermore, measurement and sampling processes are never completely reliable. Hence, achieved observations must be interpreted very carefully, especially if the reliability of such functions is relatively low. Thus, the fusion system presented in this paper is based on a Bayesian network to specifically address this problem. Therefore, information needs of the pilots on safe landing zones are determined and required perceptive capabilities are derived. Consequently, reliability estimations of the applied perceptive capabilities are incorporated. Modelling aspects of the evaluation mechanism are explained and implications of incorporated expert knowledge are set out. The feasibility of the implemented system is tested in an exemplary rescue mission, outlining the importance of incorporating automation reliability in automated decision-support systems.

Keywords: Information fusion · Bayesian networks · Decision support · Perception management · Manned unmanned teaming · Multi-UAV

1 Introduction

Landing the aircraft is one of the most challenging and dangerous tasks in aviation as pilots must perform many workload-intensive cognitive tasks simultaneously in a complex environmental situation. This is especially true for field landings of helicopters taking place in uncontrolled and unsafe areas as required e.g. during military search & rescue (SAR) operations like CASEVAC (casualty evacuation) or CSAR (combat search and rescue) [1]. In such situations, the pilots must not only cope with the landing procedure itself, but also with the vulnerability of the H/C during landing and take-off [2], making it necessary to reconnoiter the landing zone in advance.

In this scope, one of the research topics in the R&D project CASIMUS (Cognitive Automated Sensor Integrated Unmanned Mission System) investigates the usage of multiple unmanned aerial vehicles (UAVs) to provide a manned two-seated transport helicopter (H/C) with up-to-date recce and surveillance data of potential landing zones in military SAR missions. To cope with the time-criticality of the latter, the UAVs are guided from on-board the H/C by the pilot-in-command (PiC) in a manned-unmanned-teaming (MUM-T) fashion to reduce typical command & control (C2) latencies and increase operational flexibility by employing higher levels of interoperability (LOI 4/5, [3]). Figure 1 depicts this functional principle, showing the H/C cockpit and three UAVs in a CASEVAC setting.



Fig. 1. MUM-T principle in our H/C mission simulator at the IFS. The PiC (left) is commanding multiple UAVs to reconnoiter its flying route (white) and mission-critical areas, i.e. the desired landing zone in the background (red). (Color figure online)

However, shifting the C2-loop into the cockpit comes with a cost. In contrast to legacy unmanned aerial system (UAS), the PiC must handle all UAV-related tasks in addition to his conventional task spectrum. Thus, a naïve MUM-T approach bears the risk to greatly increase crews workload which needs to be monitored and balanced in some way [4], either by the crew themselves or by an associate system on-board the H/C [5–7]. A promising way for workload mitigation is to adapt the task sharing between human and machine by adopting varying levels of automation (LOA, [8, 9]). Thus, to enable higher LOA, the UAVs must be capable to perform certain tasks in a (semi-)autonomous manner. Nevertheless, employing higher LOA bears the risk of automation-induced errors (“automation surprises”), complacency effects or other Trust-in-Automation issues [10–12]. To cope with such effects the behavior of the automated systems should be pilot-understandable and self-explanatory.

Therefore, this article describes a decision support system (DSS) for the multi-UAV based reconnaissance and assessment of helicopter landing zones to aid the H/C crew in picking a safe and suitable landing point during mission. The DSS is built upon the *Perception-Oriented Cooperation Agent* (POCA) [13], using a Bayesian network approach to evaluate possible landing points while providing self-explanation capabilities through diagnostic inference.

The remainder of this article is structured as follows: Sect. 2 sums up previous and related work in automated landing zone reconnaissance and agent self-explanation mechanisms. A general system is given in Sect. 2, stating requirements and providing operational principles for multi-UAV based landing zone reconnaissance. Section 4 describes the Bayesian network approach used in the landing point evaluation and self-explanation mechanisms. Preliminary evaluation results are presented in Sect. 5 along with some integration aspects in a full mission H/C simulator. Finally, Sect. 6 concludes the article and gives an outlook to further research and planned experiments.

2 Related Work

Landing Zone Reconnaissance or Landing Site Detection is a common problem in manned and unmanned aviation as well as in space exploration. In the following, some of the surveyed articles are reflected.

In manned aviation, landing zone reconnaissance often denotes the problem of flying in degraded visual environments (DVE). Therefore, Szoboszlai et al. [14–17] investigated the usage of LIDAR technology to detect a H/C landing site under DVE conditions and integrated the visualization in the helmet-mounted-display of the H/C pilot. They conducted research on necessary symbology and proofed their system in several flight test campaigns. A similar system was developed by Airbus [18], incorporating more sophisticated means of landing site and obstacle visualization.

Likewise, systems for the detection of safe landing points are presented in the unmanned aviation domain. Fitzgerald et al. [19, 20] combined basic computer vision algorithms with neural network based texture classifiers for surface type detection to select the best LS in a single image in case of UAV emergency. Patterson et al. developed a comparable system for the same problem in [21]. However, they are only relying on the detection of free areas in a single monocular image to determine a safe landing sites by using a simple edge extraction algorithm. In [22] their concept is extended to incorporate data obtained by other UAVs or using human operator input.

In the same scope, Coombes et al. [23] used a Multi Criteria Decision Making (MCDM) Bayesian Network (BN) for landing site selection. In their approach the proposed decision-making BN selects the emergency landing site based on General Aviation (GA) requirements on emergency landing sites.

Apart from that emergency LS detection problem, much work was done by Scherer et al. [24, 25] to determine a suitable landing site for an unmanned full-size helicopter. The proposed system heavily relies on a LIDAR-created 3D point cloud to create an elevation map allowing a rough evaluation of free areas. Besides the point cloud information, various other factors are considered, including terrain clearance, approach/depart paths, and wind direction. The selection itself is based on a *goodness* function,

linearly combining the different selection criteria, incorporating operator preference in terms of weight adjustments.

Furthermore, space exploration missions demand a safe and reliable mechanism for automated landing site suitability determination during spacecraft descent. Therefore, Serrano [26] proposed a selection system based on Bayesian Networks (BN), integrated in a multi-sensor framework comprising of RADAR, LIDAR and camera sensors. Thereby, the presented decision system incorporated not only classical safety-related criteria, but also additional mission-specific factors, as for example the expected scientific return.

Our approach now incorporates sensors on multiple UAVs in the decision-making process. Therefore, the idea of modelling multiple decision criteria in a Bayesian Network [23] was picked up and extended to incorporate the perceptive reliability for landing point suitability determination.

3 Multi-UAV-Based Landing Zone Reconnaissance

Landing Zone Reconnaissance (LZR) denotes the task of reconnoitering a designated area (landing zone, LZ) to examine its suitability for take down, incorporating possible threats and physical characteristics of the landing zone. In this regard, performing LZR for a manned H/C in a full-fledged military rescue mission differs from the approaches presented before (cf. Sect. 2) as additional tactical and mission-critical aspects must be considered, most importantly the reliability of highly automated perceptive subfunctions [27, 28].

Figure 2 depicts an example setup for such a CASEVAC mission. There, a manned transport helicopter supported by three UAVs is deployed to rescue a group of persons in an unsafe operation area, whereby their last known position determines the designated landing zone. In this MUM-T setup, the UAVs shall reconnoiter potential landing points for the H/C suitable for a successful evacuation. Therefore, the UAVs must gather numerous and heterogeneous data from the multiple potential landing points and evaluate them accordingly.

In the following, landing zone selection and landing point evaluation criteria are stated. Subsequently, the general concept for multi-UAV based LZR is presented.

3.1 Landing Zone Selection Criteria

Different regulations or heuristics exist for the definition of safe landing zones, both in civilian and in military applications. Basic requirements for conducting military LZR are stated in [1], leaving much space for national implementation, as for example the publicly available LZR regulations by the U.S. Army [29]. In general, several heterogeneous information needs must be incorporated when reconnoitering a landing zone: tactical, aeronautical, and meteorological. These needs come with inherent sequential ordering - for example, flight safety related considerations as obstacle situation can be neglected if tactical clearance has already failed.

Below, the current regulatory situation is summarized. Additional information gathered in consultative talks with German army aviators is incorporated.

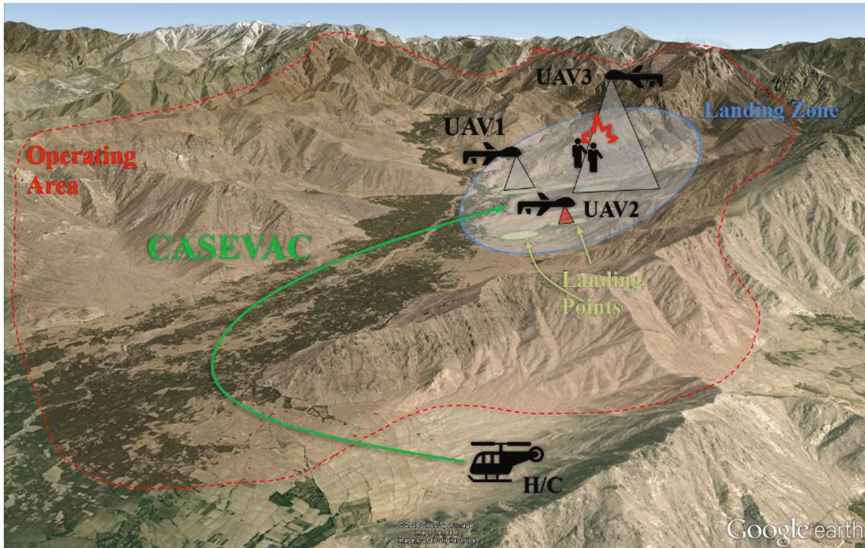


Fig. 2. CASEVAC example scenario: a manned transport helicopter aims in rescuing persons in an unsafe environment, guiding a team of three UAVs in a MUM-T fashion to determine the most suitable landing point in the designated landing zone (light blue). (Color figure online)

Tactical Considerations

The dominant concern for military LZR is information on the tactical situation and on mission-critical observations:

- **Safety.** Is the LZ safe or do threats exist? Are buried objects as EODs or mines present in the LZ? Are streets nearby, thus allowing enemy forces to reach the LZ fast?
- **Mission achievability.** Is the mission achievable from the LZ? How long will it take to reach the mission objective from the LZ? How long will the H/C be grounded and thus remain vulnerable?

Aeronautical Considerations

This contains requirements related to flight safety aspects in the LZ:

- **Landing zone size.** Is enough space for the actual H/C to touch down available?
- **Approach and departure directions.** Are obstructions in the approach or depart vectors present? Are conditions given limiting the approach/depart directions? What if the H/C is fully-loaded?
- **Obstacle situation.** Is the landing point itself free of obstacles? Smaller obstacles (debris, < 0.45 m) can be ignored;
- **Ground slope.** Does the slope exceed the H/C’s limits? (However, the pilots can hover if ground slope is too high.)
- **Surface type and conditions.** What kind of surface will be encountered? Is there a risk to bog down? Are there brown-out or white-out conditions?

- **Ground solidity and load suitability.** Can a heavy transport H/C touch down on the desired landing point? Is the landing gear of the H/C suitable for landing in the desired area?

Meteorological Considerations

This refers to meteorological conditions in the LZ, directly influencing the H/Cs capability or risk for landing:

- **Cloud ceiling and visibility.** What is the ceiling level? Is it raining or do we have fog?
- **Density altitude.** What is the comparable density altitude at the LZ? Thus, is the H/C performant enough to operate in the LZ, even under high-load conditions?
- **Wind conditions.** What are wind velocities and directions? Thus, is landing into the wind possible? Do crosswind or tailwind conditions prevail?

3.2 Multi-UAV Concept

As depicted in Fig. 2 and stated in the prior section, landing zone reconnaissance requires the gathering and evaluation of numerous and heterogeneous data from (multiple) potential landing points in a designated area, the landing zone. Thus, the UAVs must provide a broad range of perceptive capabilities. However, UAVs are often restricted in terms of sensor payload capacity and on-board processing resources, effectively resulting in a limited set of capabilities. Hence, a single UAV might be insufficient to satisfy the perceptive requirements for a complex task as LZR. A promising way to overcome these limitations is the teaming of multiple UAVs by combining their capabilities in a cooperative manner, profiting from heterogeneous payload setups and varying platform characteristics as well as from task parallelization opportunities.

Consequently, we proposed the system concept of the Perception Oriented Cooperation Agent (POCA) in [13], extending the Sensor- & Perception Management (SPM) paradigm described in [30]. Thereby, it integrates environmental and platform self-adaption mechanisms from the SPM system while additionally incorporating perception planning and scheduling capabilities allowing to benefit from the multi-UAV set up stated above.

Figure 3 depicts the basic system concept. Here the operator (Pilot in Command) issues a “Landing Zone Recce” task to the system in a supervisory control manner, along with external constraints as the landing zone boundaries and available resources, e.g. available UAVs and thus available sensory equipment. This task is then analyzed by POCA to extract required perceptive actions, thus reflecting the information needs described in Sect. 3.1. These subtasks are interpreted as primary planning goals for the integrated task planning and scheduling mechanisms. During planning, external constraints, specific task requirements and available UAV capabilities extracted from a Perception Resource and Capability Ontology [31] are incorporated. Combining these, the Perception Planner creates a task agenda comprised of interleaved perceptual and navigational subtasks which are subsequently used to control and coordinate the UAVs underlying automation functions, i.e. the UAVs flight management system (FMS) and the SPMS. Thereby, plan generation itself follows a classical team-leader/team-member

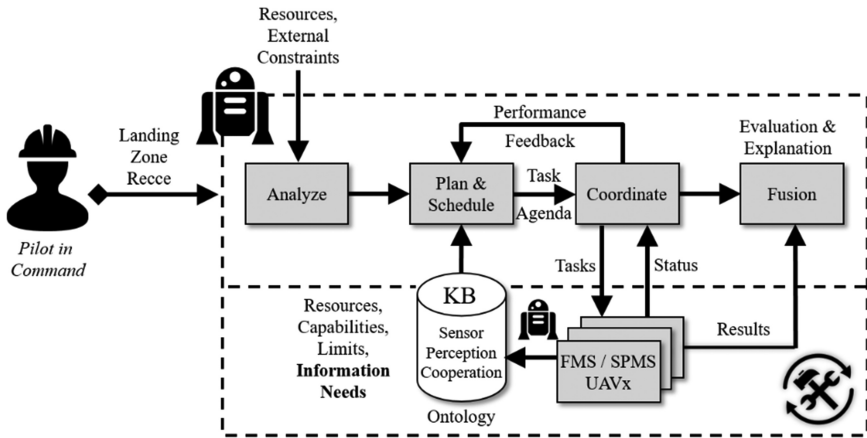


Fig. 3. Operational principle of POCA [13]. The notation is based on the work system notation in [33], thus the supervisory control arrow denotes both control and information feedback flow.

structure: the plan itself is generated by a designated team leading UAV whereas each team member is responsible for execution of the subtasks. During execution, the single POCA instances onboard the UAVs gather the results of their scheduled perception tasks, e.g. obstacle or vehicle detection. The results are transmitted to the leading UAV, which fuses and assesses the gathered results to derive a recommendation on the best suited landing zone.

This article focus on this latter step. In the following we describe an information-level fusion agent [32] for landing point evaluation based on causal Bayesian Inference, thereby incorporating perceptive reliability of the automated reconnaissance functions and expert knowledge on the underlying perceptive subtasks as well as their interconnections.

4 Bayesian Landing Point Evaluation

The fusion of heterogeneous information reflecting different physical phenomena, as needed for landing zone reconnaissance (cf. Sect. 3.1), invalidates the usage of classical low-level fusion methods such as Kalman filtering. Thus, a more abstract representation for such incommensurate data is needed to enable higher level fusion mechanisms on information level [32]. Hence, preprocessing and pre-assessment of the underlying sensory data is mandatory to enable the subsequent fusion mechanisms.

In POCA such preprocessing is realized by incorporating the SPM system of Hellert and Smirnov [34] (cf. Fig. 3), allowing it to rely on the integrated perceptive capabilities and high-level inference mechanisms to obtain semantically enriched results. Consequently, in POCA information-level fusion is applied on the percepts retrieved from the underlying SPM instances.

However, automated perception functions are imperfect by design and must be handled carefully, as such exhibit non-deterministic behavior and are prone to inherent uncertainties due to implementation weaknesses or changing operational

environments [27, 28]. Thus, the outcome and results of the perceptive subtasks can best be expressed probabilistically. Therefore, to safely assess landing points in an examined landing zone a fusion mechanism is needed capable of handling these uncertainties.

In addition, the fusion architecture shall be able to incorporate expert knowledge on the information needs and provide means for extension and modularization allowing to adapt the fusion architecture on new or changed applications, e.g. civilian search & rescue missions.

Considering the above, we propose the usage of a *Bayesian Network* (BN) [35] to explicitly model knowledge on the interdependencies between the information needs in a fusion graph. Thereby, the conditional probabilities of the network are automatically adjusted during runtime.

In the following, some BN fundamentals are stated. Afterwards, our approach to evaluate landing points using a BN is explained in Subsect. 4.2.

4.1 Bayesian Network Fundamentals

BNs are a commonly used graphical tool for knowledge representation and reasoning under uncertainty in decision-making intelligent systems, allowing the incorporation of explicit modelled and elicited knowledge of domain experts.

More generally, a BN is a *Directed Acyclic Graph* (DAG) in which the nodes represent the random variables of interest and the arcs the causal relation between these nodes, thus reflecting the conditional dependency between the nodes. In addition, a BN assumes conditional independency between nodes on the same level, meaning that any node x_i with the parents y_i is conditionally independent from any other variable except of its descendants z_i . Thereby, the graphical representation of BNs provides an unambiguous and relatively simple way of representing this independency between variables.

Figure 4 visualizes this independency topology and shows the three fundamental connection types for nodes in a BN, forming the basic conditional probabilities for BNs. Thus, a joint probability distribution (JPD) for a BN with the nodes $X_i = \{x_i, \dots, x_n\}$ can be derived using Bayes' chain rule:

$$P(x_1, \dots, x_n) = P(x_n | x_{n-1} \dots x_1) P(x_{n-1} | x_{n-2} \dots x_1) \dots P(x_2 | x_1) P(x_1) \quad (1)$$

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | x_{i-1}, \dots, x_1) \quad (2)$$

Incorporating the independency assumption above with the parents $Y_i = \{y_i, \dots, y_n\}$, Eq. (2) could be simplified to:

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | y_i(X_i)) \quad (3)$$

This simplified JPD exhibits an important feature of BNs: since the node x_i in a BN is only dependent on the state of its parents y_i instead of depending on arcs to each other node $x_j \in X_i$ (which requires 2^n arcs), the number of parameters needed to model or learn a BN can be reduced drastically.

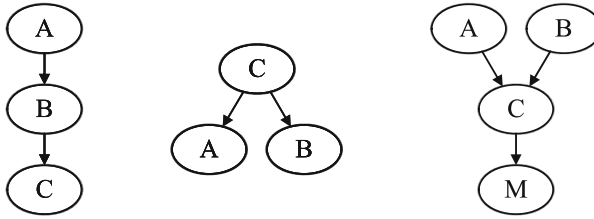


Fig. 4. Basic elements of BNs: serial connection (left), diverging connection (middle) and converging connection (right).

With *Influence Diagrams* (ID) a generalization for BNs exists to allow the application on decision-making problems [36]. IDs add two special types of nodes in the BN notation (*decision* and *utility* nodes), whereas random variables are called *chance* nodes. Thereby, a decision node is a controllable point where a mutual exclusive action $A = \{a_1, \dots, a_n\}$ influences the probability distributions off connected random variables. Utility nodes represent the value or outcome of a decision. In multi-criteria decision-making (MCDM) problems, criteria nodes denote chance nodes directly influencing a utility node [37].

4.2 Landing Point Evaluation Using a Bayesian Network

As described earlier, various criteria are needed to safely assess and evaluate a landing zone and to pick a safe and reliable landing point. Thus, it comes naturally to formulate the landing point evaluation problem in terms of multi-criteria decision-making (MCDM).

In the following our approach for a MCDM Bayesian Network to evaluate landing points, following the notations in [37]. The single components of the network are explained, providing insights on implementation details. Figure 5 visualizes the DAG of the developed BN while Table 1 lists all nodes with their possible, discretized states and associated node types.

Essentially, the requirements in Sect. 3.1 can be summarized in two criteria: helicopter safety and mission achievability. Consequently, the utility end-node “Landing Point Quality” in Fig. 5 is only influenced by the two reflecting criteria nodes “Landing Point Safety” and “Mission Achievability”, which combine the conditional probabilities tables (CPT) of the underlying information needs encoded in the single chance nodes. Effectively, the utility node implements a weighting function of the two mission-influencing criteria, thereby prioritizing helicopter safety.

In our current implementation, mission achievability is only influenced by the geographical distance to the mission objective (e.g. the distance to the last known position of the persons to be rescued in the example in Fig. 2). In contrast, helicopter safety is influenced by a broad variety of parameters.

Expert knowledge on the interconnections between information needs is encapsulated semantically in the structure of the net in Fig. 5 and the CPT of the criteria nodes. Thus, to ease the compilation of the CPT for the “Landing Point Safety” node, several intermediate or hidden nodes were used.

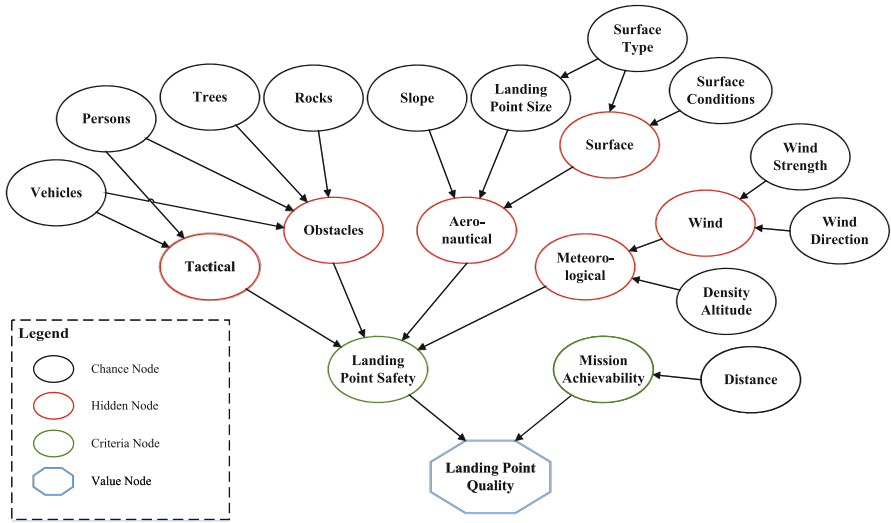


Fig. 5. DAG structure for the proposed landing point evaluation network.

Table 1. List of all nodes, their discrete states and the associated node type in the proposed landing point evaluation network. Nodes are sorted according their related type.

Node	States	Node type
Vehicles	Present; Absent	Chance
Persons	Present; Absent	Chance
Trees	Present; Absent	Chance
Rocks	Present; Absent	Chance
Slope	Very High; High; Medium; Low; None	Chance
Landing point size	Tiny; Small; Medium; Big	Chance
Surface type	Grass; Concrete; Swamp; Sand; Snow	Chance
Surface conditions	Dry; Wet	Chance
Wind strength	Strong; Medium; Weak; None	Chance
Wind direction	Head; Cross; Tail	Chance
Density altitude	Comparable; Different	Chance
Distance	Close; Medium; Out of Range	Chance
Tactical	Safe; Unsafe	Hidden
Obstacles	Present; Absent	Hidden
Surface	Good; Bad	Hidden
Aeronautical	Good; Bad	Hidden
Meteorological	Safe; Unsafe; Dangerous	Hidden
Safe landing point	Safe; Unsafe	Criteria
Mission achievability	Possible; Critical; Impossible	Criteria
Landing point quality	Quality percentage (0–100)	Value

The chance nodes represent the sensed input to the proposed fusion system. CPT values are expressed probabilistically to reflect inaccurate measurements and nondeterministic behavior of the underlying processing algorithms. Quantitative values for the CPTs are constantly updated during mission, thereby incorporating potentially varying automation reliabilities for the actually selected perceptive functions, which can be adjusted during mission due to environmental changes [28]. Perceptive tasks with quantifiable results, for example vehicle or obstacle detection are heavily condensed (cf. Table 1). Thus, the insignificance for the safety assessment is expressed whether there are one or ten potentially dangerous objects at a landing point. Continuously valued results, e.g. the slope in degrees, are discretized accordingly to enable incorporation in the BN structure [38]. In addition, the influence of the helicopter type on the aeronautical and meteorological criteria is explicitly modelled in the BN, using a decision node. However, for the sake of greater clarity this is neither depicted in Fig. 5 nor listed in Table 1.

Finally, causal reasoning is applied using the SMILE engine¹ [39] to gain an actual quality value for the currently processed landing point.

5 Results and Discussion

To demonstrate the feasibility of the presented landing point evaluation and fusion mechanism, an example use case scenario for landing zone reconnaissance was created. The tactical situation for the unreconnoitered landing zone in the test setup is depicted in Fig. 6, embedded in the bigger scope of a full CASEVAC mission outlined before (cf. Sect. 3). There, a group of persons must be rescued in an unsecure and potentially dangerous area, whereby their last position is known, thus determining the rescue area and the designated landing zone. The rescuing H/C is supported by a team of three UAVs, providing perceptive capabilities for landing zone reconnaissance.

The parameters for each chance node in the test scenario are shown in Table 2. To reflect state crossings, a soft threshold was used to determine the discrete states. For example, the surface type of LPB is not clearly determinable. To account for such ambiguities, the CPTs of the continuous attributes may contain factorized values. Thus, the surface type of LPB is set to 85% of grassland and 15% of sand, reflecting an area with a rough grass cover containing some sandy spots.

Not depicted in Table 2 are the probabilistic influences of the used perceptive capabilities. The values for the appropriate measurements are applied to the single chance nodes individually. We used reliability values for perceptive algorithms available in our SPM systems [30, 34]:

- Person detection was performed using an infrared camera based support vector machine (SVM) classifier, having a relatively low reliability of 0.66 [27].
- For vehicle detection a deformable part model (DPM) with a trained SVM classifier on electro-optical images was deployed, having a reliability of 0.93 [28].

¹ BayesFusion, LLC, <http://www.bayesfusion.com/>.

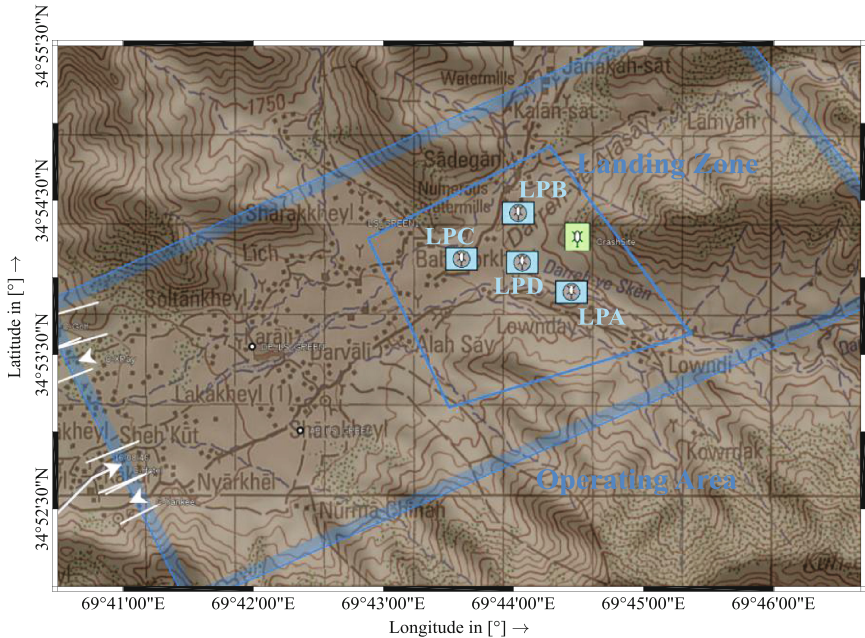


Fig. 6. Test setup for landing zone reconnaissance in a CASEVAC scenario. The landing zone contains four landing points LPA, LPB, LPC and LPD to be reconnoitered and evaluated.

Table 2. Discrete parameters of the chance nodes for the single landing points. A soft threshold was applied, thus factorized values reflect transition between the discrete states.

	LPA	LPB	LPC	LPD
Vehicles	Present	Absent	Absent	Absent
Persons	Absent	Absent	Absent	Absent
Trees	Present	Absent	Present	Absent
Rocks	Absent	Absent	Absent	Present
Slope	Low	Low	Low	Medium
LP size	Medium	Big	Big	Small
Surface type	0.5 Grass 0.5 Sand	0.85 Grass 0.15 Sand	0.3 Grass 0.7 Sand	Grass
Surface conditions	Dry	Dry	Dry	Dry
Wind strength	Medium	Weak	Medium	Weak
Wind direction	0.3 Cross 0.7 Tail	0.8 Head 0.2 Cross	0.3 Head 0.7 Cross	Head
Density altitude	Comparable	Comparable	Comparable	Comparable
Distance	Close	Close	Medium	Close

- Rock and tree detection as well as landing point and slope measurement are based on LIDAR processing, having a measurement and detection reliability of 0.98.
- Surface determination is based on GIS data for which a deterministic value is assumed. Nevertheless, the nodes incorporate soft thresholding as described above.
- The same applies for the meteorological nodes, incorporating data from a weather information service.
- Distance measurement is based on a simple estimation of the walking distance, combining Euclidean distance and the movement speed of persons by foot.

Table 3. Evaluation results and quality values for the single landing points.

	Safety	Achievability	Quality
LPA	Unsafe: 7%	Possible: 100%	69%
LPB	Safe: 60%	Possible: 100%	59.9%
LPC	Unsafe: 21%	Critical: 50%	18.1%
LPD	Unsafe: 18%	Possible: 100%	17.5%

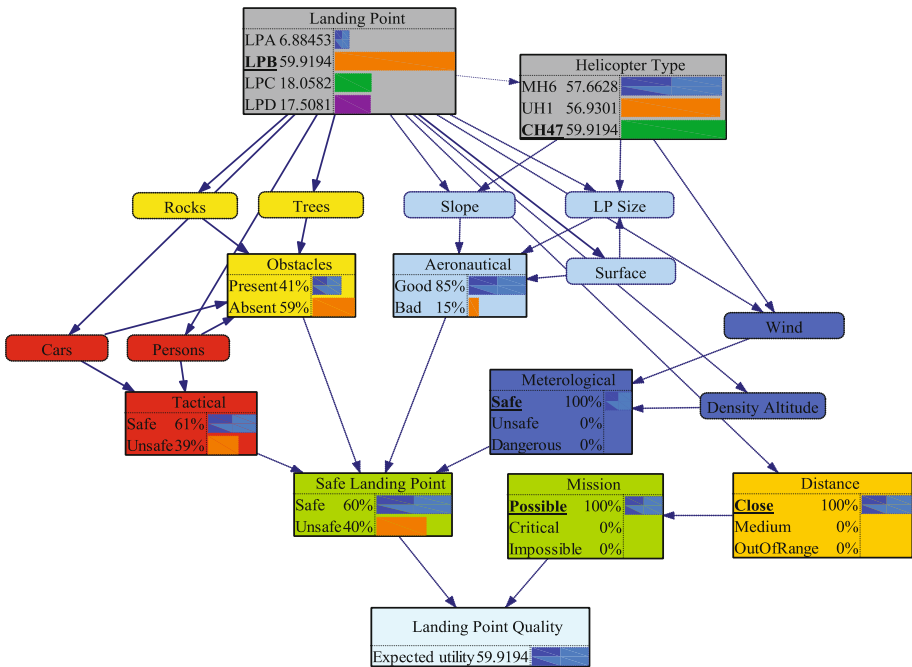


Fig. 7. Results of the landing point evaluation for LPB, modelled in GeNIe [36]. Chance nodes belonging in the same group are color-coded: (tactical), yellow (obstacle), light blue (aeronautical) and dark blue (meteorological). Submodels (e.g. “Cars” or “Persons”) were used to apply the automation reliabilities on the CPTs of the appropriate chance nodes. Due to modelling and test purposes in GeNIe, an additional decision node (“Landing Points”, gray) exist. (Color figure online)

The quality of each landing point was determined assuming a Boeing CH-47 as rescuing helicopter. The criteria values and estimated quality inferred are summarized in Table 3. The network selects LPB as recommended landing point for the rescue mission as it has the highest quality assessment of 59.9%. Although this value seems to be rather low, it characterizes the importance of incorporating automation reliability in the decision-making process when applying highly-automated sensor based systems.

Figure 7 displays the Bayesian network modelled in GeNIe [39] with inferred values for the winning landing point LPB. As it can be seen, the quality assessment is mainly based on the safety estimation. Thus, the usage of the relatively unreliable person detector described above heavily influences the quality estimation, following Bayes rule for determining the JPD as shown in Eq. (3).

A naïve interpretation of the parameters in Table 2, ignoring automation reliability, might have led to the false impression of safety. Consequently, a decision-support mechanism for highly automated reconnaissance systems must consider possible drawbacks when presenting results and providing suggestions to human operators.

Thus, the presentation of the evaluation results and final recommendation shall express the uncertain nature of the decision-making process and allow the pilot to scrutinize the derived result. Our current approach for presenting the evaluation results on recommendation level uses a color-coded traffic-light representation as shown in Fig. 8. Whenever the pilot chooses to receive more details on the decisions rational, information on the most influencing factors are presented (not depicted).



Fig. 8. Tactical map visualization of the landing point evaluation results displayed in the multi-function-display of our H/C simulator. The landing point quality is depicted color-coded in an easy-to-understand traffic-light representation.

6 Conclusion

The safe and reliable reconnaissance of a helicopter landing zone requires the gathering of various heterogeneous data from potential landing points which must be fused and evaluated accordingly. Therefore, such an evaluation and fusion mechanism was presented in this paper based on a multi-criteria decision-making Bayesian Network. The proposed BN incorporates expert knowledge on LZR information needs and a probabilistic representation of the automation reliability, adapted online during mission. Bayesian inference is applied to estimate the landing point quality whenever new reconnaissance data comes available.

The feasibility of presented fusion agent was demonstrated on a given example and obtained results are demonstrated and discussed, emphasizing the importance of incorporating automation reliability in the decision-making process. Probabilistic inclusion of the reliability value in a Bayesian Network as presented here provides a promising way to deal with such influences and resulting automation over trust issues. An easily understandable visualization concept for the evaluation results in the multi-function-display of a H/C cockpit is presented.

Nevertheless, to avoid distrust effects, further work is required to increase the overall system reliability. For example, the automated system can apply additional fusion mechanisms to fulfill a perceptive requirement more reliably, e.g. by combining multiple perception algorithms in the person detection processing chain. Another promising approach is to incorporate inputs of a human operator in cases when the overall system reliability is too low to be trusted [7].

Furthermore, additional work is required on the result presentation to enable the crew to verify the reconnaissance results by themselves and thus to better understand the landing point recommendation.

Next steps will quantifiably determine benefits and overall system acceptance when interacting with military trained H/C pilots. Therefore, an experimental operator-in-the-loop campaign in the full mission MUM-T H/C simulator are planned for summer 2017. In addition, technological readiness for multi-UAV based perception will be evaluated in a down-sized Landing Zone Reconnaissance experiment at university grounds in summer too.

References

1. NATO Standardization Agency: STANAG 2999 - Use of Helicopters in Land Operations Doctrine ATP-49(F) (2012)
2. Hastert, P.L.: Operation Anaconda: perception meets reality in the hills of Afghanistan. *Stud. Confl. Terror.* **28**, 11–20 (2005)
3. NATO Standardization Agency: STANAG 4586 - Standard Interface of UAV Control System (UCS) for NATO UAV Interoperability (2012)
4. Schulte, A., Donath, D., Honecker, F.: Human-system interaction analysis for military Multi-RPA pilot activity and mental workload determination. In: *IEEE International Conference on Systems, Man, and Cybernetics (SMC 2015)*, pp. 1375–1380 (2015)

5. Strenzke, R., Uhrmann, J., Benzler, A., Maiwald, F., Rauschert, A., Schulte, A.: Managing cockpit crew excess task load in military manned-unmanned teaming missions by dual-mode cognitive automation approaches. In: AIAA Guidance, Navigation, and Control Conference (2011)
6. Honecker, F., Brand, Y., Schulte, A.: A task-centered approach for workload-adaptive pilot associate systems. In: The 32nd EAAP Conference, Cascais, Portugal (2016)
7. Ruf, C., Stütz, P.: Model-driven sensor operation assistance for a transport helicopter crew in manned-unmanned teaming missions: selecting the automation level by machine decision-making. In: Savage-Knepshield, P., Chen, J. (eds.) *Advances in Human Factors in Robots and Unmanned Systems*. AISC, vol. 499, pp. 253–265. Springer, Cham (2017). doi:[10.1007/978-3-319-41959-6_21](https://doi.org/10.1007/978-3-319-41959-6_21)
8. 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**, 286–297 (2000)
9. Sheridan, T.B.: Adaptive automation, level of automation, allocation authority, supervisory control, and adaptive control: distinctions and modes of adaptation. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **41**, 662–667 (2011)
10. Sarter, N.B., Woods, D.D., Billings, C.E.: Automation surprises. In: Salvendy, G. (ed.) *Handbook of Human Factors and Ergonomics*, pp. 1926–1943. Wiley (1997)
11. Parasuraman, R., Molloy, R., Singh, I.L.: Performance consequences of automation-induced “Complacency”. *Int. J. Aviat. Psychol.* **3**, 1–23 (1993)
12. Baker, A.L., Keebler, J.R.: Factors affecting performance of human-automation teams. In: Savage-Knepshield, P., Chen, J. (eds.) *Advances in Human Factors in Robots and Unmanned Systems*. AISC, vol. 499, pp. 331–340. Springer, Cham (2017). doi:[10.1007/978-3-319-41959-6_27](https://doi.org/10.1007/978-3-319-41959-6_27)
13. Schmitt, M., Stuetz, P.: Perception-oriented cooperation for multiple UAVs in a perception management framework: system concept and first results. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, California, USA, pp. 1–10. IEEE (2016)
14. Szoboszlay, Z.P., Turpin, T.S., McKinley, R.A.: Symbology for brown-out landings: the first simulation for the 3D-LZ program. In: 65th American Helicopter Society Annual Forum 2009 (AHS65). American Helicopter Society International (AHS), Ft. Worth, Texas, USA (2009)
15. Szoboszlay, Z.P., McKinley, R.A., Braddom, S.R., Harrington, W.W., Burns, H.N., Savage, J.C.: Landing an H-60 helicopter in brownout conditions using 3D-LZ displays. In: 66th American Helicopter Society Annual Forum 2010 (AHS66). American Helicopter Society International (AHS), Phoenix, Arizona, USA (2010)
16. Harrington, W., Braddom, S., Savage, J., Szoboszlay, Z., McKinley, R.A., Burns, H.N.: 3D-LZ Brownout landing solution. In: 66th American Helicopter Society Annual Forum 2010 (AHS66). American Helicopter Society International (AHS), Phoenix, Arizona, USA (2010)
17. Szoboszlay, Z.P., Fujizawa, B.T., Ott, C.R., Savage, J.C., Goodrich, S.M., McKinley, R.A., Soukup, J.R.: 3D-LZ flight test of 2013: Landing an EH-60L Helicopter in a brownout degraded visual environment. In: 70th American Helicopter Society Annual Forum 2014 (AHS70). American Helicopter Society International (AHS), Montréal, Québec, Canada (2014)
18. Airbus Defence & Space: White Paper: Solving the problem of flying in DVE for helicopter pilots (2014)

19. Fitzgerald, D., Walker, R.: Classification of candidate landing sites for UAV forced landings. In: AIAA Guidance, Navigation, and Control Conference and Exhibit. American Institute of Aeronautics and Astronautics, San Francisco, California, USA (2005)
20. Fitzgerald, D., Walker, R., Campbell, D.: A Vision based forced landing site selection system for an autonomous UAV. In: 2005 International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Melbourne, Australia, pp. 397–402. IEEE (2005)
21. Patterson, T., McClean, S., Parr, G., Morrow, P., Teacy, L., Nie, J.: Integration of terrain image sensing with UAV safety management protocols. In: Par, G., Morrow, P. (eds.) *Sensor Systems and Software*, pp. 36–51. Springer, Heidelberg (2011)
22. Patterson, T., McClean, S., Morrow, P., Parr, G.: Modelling safe landing zone detection options to assist in safety critical UAV decision making. *Procedia Comput. Sci.* **10**, 1146–1151 (2012)
23. Coombes, M., Chen, W.-H., Render, P.: Site selection during unmanned aerial system forced landings using decision-making Bayesian networks. *J. Aerosp. Inf. Syst.* **13**, 491–495 (2016)
24. Scherer, S., Chamberlain, L., Singh, S.: Online assessment of landing sites. In: AIAA Infotech@aerosp., pp. 1–14 (2010)
25. Scherer, S., Chamberlain, L., Singh, S.: Autonomous landing at unprepared sites by a full-scale helicopter. *Rob. Auton. Syst.* **60**, 1545–1562 (2012)
26. Serrano, N.: A Bayesian framework for landing site selection during autonomous spacecraft descent. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, pp. 5112–5117. IEEE (2006)
27. Russ, M., Stuetz, P.: Application of a probabilistic market-based approach in UAV sensor & perception management. In: 2013 16th International Conference on Information Fusion (FUSION 2013), Istanbul, pp. 676–683 (2013)
28. Hellert, C., Stütz, P.: Performance prediction and selection of aerial perception functions during UAV missions. In: 2017 IEEE Aerospace Conference, Big Sky, Montana. IEEE (2017, in press)
29. U.S. Department of the Army: Attack Reconnaissance Helicopter Operations (FM 3-04.126) (2007). http://usacac.army.mil/sites/default/files/misc/doctrine/CDG/cdg_resources/manuals/fm3_04x126.pdf
30. Russ, M., Stütz, P.: Airborne sensor and perception management: a conceptual approach for surveillance UAS. In: Proceedings of the 15th International Conference on Information Fusion (FUSION 2012), Singapore, pp. 2444–2451. IEEE (2012)
31. Smirnov, D., Stütz, P.: Use case driven approach for ontology-based modeling of reconnaissance resources on-board UAVs using OWL. In: 2017 IEEE Aerospace Conference, Big Sky, Montana, USA. IEEE (2017, in press)
32. Hall, D.L., Llinas, J.: Multisensor data fusion. In: *Handbook of Multisensor Data Fusion*, pp. 1–14. CRC Press (2008)
33. Schulte, A., Donath, D., Lange, D.S.: Design patterns for human-cognitive agent teaming. In: Harris, D. (ed.) *EPCE 2016. LNCS (LNAI)*, vol. 9736, pp. 231–243. Springer, Cham (2016). doi:10.1007/978-3-319-40030-3_24
34. Hellert, C., Smirnov, D., Russ, M., Stuetz, P.: A high level active perception concept for UAV mission scenarios. In: *Deutscher Luft- und Raumfahrtkongress 2012*, pp. 1–9. Deutsche Gesellschaft für Luft- und Raumfahrt - Lilienthal-Oberth e.V., Berlin (2012)
35. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo (1988)
36. Russell, S.J., Norvig, P.: making simple decisions. In: *Artificial Intelligence: A Modern Approach*, 2nd edn., pp. 626–627. Prentice Hall (2010)

37. Watthayu, W., Peng, Y.: A Bayesian network based framework for multi-criteria decision making. In: Proceedings of the 17th International Conference on Multiple Criteria Decision Analysis, Whistler, B.C., Canada (2004)
38. Russell, S.J., Norvig, P.: Probabilistic Reasoning. In: Artificial Intelligence: A Modern Approach, 3rd edn., pp. 510–565. Prentice Hall (2010)
39. Druzdel, M.J.: SMILE: structural modeling, inference, and learning engine and GeNIe: a development environment for graphical decision-theoretic models. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence (AAAI 1999/IAAI 1999), pp. 902–903. AAAI Press, Menlo Park (1999)

Factors Influencing Cargo Pilots' Fatigue

Rui-shan Sun^(✉), Zi-li Chen, Guang-xia Huang-fu, Guang-fu Ma,
Di Wu, and Zhen Liu

Civil Aviation University of China, Tianjin, China
sunrsh@hotmail.com, zili2017@hotmail.com

Abstract. In recent years, cargo pilots' training and reserves in China have been unable to meet the needs of the development of cargo aviation, due to the rapid development of the country's cargo aviation and the continuing increase in air traffic. Frequent night and cross-time-zone flights in particular worsen the severity of pilots' fatigue situation. At present, cargo pilots' fatigue is a key problem that affects the development of China's cargo aviation. In this study, a survey was carried out with Cargo Pilot's Fatigue Survey Scale, and the factors influencing cargo pilots' fatigue were analyzed and compared with the fatigue of the airline pilots. The results indicated that: compared with the airline pilots, cargo pilots have a higher degree of fatigue and are more likely to be conscientious and open; we also determined that number of night flights, workload, and health history are highly correlated with pilots' degree with fatigue. Nevertheless, no significant difference in workload, age, or sleep quality was noted between cargo pilots' and airline pilots' fatigue.

Keywords: Cargo pilots · Fatigue · Pilot fatigue survey scale · Grey correlation

1 Introduction

On August 14, 2013 at 5:00 am, an A300-600 cargo plane belonging to the United Parcel Service crashed near Birmingham International Airport in Birmingham, Alabama, killing two pilots. The National Transportation Safety Board released the accident investigation report on September 9, 2014, which pointed out that pilot's errors and fatigued piloting were the main causes of the accident. Notably, pilot fatigue has become one of the main risk factors affecting flight safety. A survey on the flight time limits conducted by the British Air Line Pilots Association, showed that 56% of pilots reported falling asleep at the cockpit and 29% recalled waking up to find another pilot asleep [1].

Several scholars have conducted in-depth research on the causes of flight fatigue. For example, Rosekind [2] investigated the impact of flight missions on pilots' sleep quality, circadian clock changes, and subjective feelings about fatigue, and determined that the main causes of fatigue were sleep deprivation and circadian rhythm changes. Similarly, Blakey [3] pointed out that pilot fatigue is the main causes of airplane crashes; in addition to work load, this fatigue results from lack of sleep, individual physiological clock disruption, abnormal scheduling, poor sleep quality, and drug use. Blakey added that people in a state of fatigue increase the risks associated with the task they are performing; therefore, crew fatigue conditions should be confirmed before a

flight to determine the ability of the individuals to complete flight tasks. Moreover, Caldwell [4, 5] suggested that long and short route pilots commonly attribute fatigue to overnight flights, jet lag, early wake-up times, time pressure, multiple flights, and long shifts. The most common causes of fatigue are lack of sleep, diurnal rhythm changes, poor sleep quality, stress, and excessive workload. SAFO [6] indicated that sleep deprivation and a heavy workload can trigger short-range pilots' fatigue; by contrast, long-range pilots' fatigue is attributable to a lack of sleep and the circadian rhythm disruption caused by cross-time-zone flights. Han [7] argued that amount of sleep, work load, and mood contribute to flight fatigue, and Ge [8] explored the relationship between flight fatigue and age in pilots who work long-haul flights. Li [9] suggested that a lack of sleep and changes in circadian rhythm are directly related to fatigue; he added that late work hours and monotonous tasks also lead to fatigue, while a lack of sleep can produce physiological and psychological decline and subsequent decline in work efficiency. He concluded that a decreased efficiency in flight tasks was likely to produce human error, and thus lead to accidents.

The CCAR-121 clearly stipulates that all-cargo transport aircraft (which can operate with a maximum load of more than 3400 kg), including large public air transport carriers, must comply with the provisions of Chapter P, which discuss pilot crew duties related to period limitations, flight time restrictions, and rest requirements. Although China's rules for pilot rest do not differentiate between airline and cargo pilots, Chinese pilot fatigue management only considers the lengths of time for work and rest periods; current Chinese rules do not effectively consider the question of cargo pilots' night flights.

In short, fatigue is a critical factor affecting the safety of cargo flights. Pilot fatigue can lead to a decline in operative capabilities, ability to judge errors, and hallucinations during flight, and can lead to serious and tragic flight accidents. Therefore, in this study we conducted a survey to analyze cargo pilot fatigue in a Chinese cargo airline. Based on the collected data, the grey correlation method was used to identify the factors that influence pilot fatigue.

2 Cargo Pilot Fatigue Questionnaire and the Grey Correlation Analysis Method

2.1 Cargo Pilot Fatigue Questionnaire

The questionnaire comprised six separate surveys: personal information survey, fatigue perception survey, work-related factors survey, sleep quality survey, life event factors survey, and personality characteristics survey. The six sections were later combined to investigate the overall fatigue statuses of pilots.

(1) Personal Information Survey

The first survey collected pilots' basic personal information, and consisted of three parts: basic information, family status, and health status. Basic information included age, total flight hours logged, typical route types and crew position; family status included marital status and number of children; and health status included dietary

habits, exercise routines, prescribed medications, and the presence or absence of any chronic disease.

(2) Fatigue Perception Survey

The second survey was based on the Fatigue Self-Rating Survey Scale developed by Prof. Tianfang Wang [10] and the MFI-20 Scale [11] developed by the psychology department at the University of Strand. By combining these surveys with questions that targeted the specific circumstances of China's cargo pilots, a new fatigue rating scale (MFI-16) suitable for pilots was developed. There are four dimensions to the survey, namely general fatigue, physical fatigue, mental fatigue, and reduced motivation, comprising a total of 16 items. The score for each dimension ranged from 4 to 20 points; all four dimensions had a total score of 16–80 points, with higher scores indicating a greater perceived degree of fatigue. This survey determined the fatigue statuses of the pilots over the course of 1 month, and demonstrated good reliability and validity.

(3) Work-Related Factors Survey

The work-related factors survey consisted of 19 items organized into four dimensions: scheduling factors, work load, work environment, and other factors.

(4) Sleep Quality Survey

The sleep quality survey adopted the content of the Pittsburgh Sleep Quality Index [12]. The scale was developed by a sleep specialist, Buysse Dj, in 1993 while he worked for the Center for Sleep and Biorhythmics at the University of Pittsburgh Medical Center to assess subjects' subjective perceptions of sleep quality over a period of 1 month. The reliability and validity of this scale have been verified by Xianchen Liu in China. Thus, it has become a common scale for studying sleep disorders and clinical evaluation.

(5) Life Event Factors Survey

The life events factor survey reviewed 17 common life events, which were organized into the following seven dimensions: workload, career development, interpersonal relationships, marriage and family, property economy, physiological status, and institutional pressure. Notably, this survey not only explored general life events that can produce considerable fluctuations in human emotions, but also occupation-related special events specific to pilots.

(6) Personality Characteristics Survey

The personality trait theory defines a "trait" as a basic characteristic of individual behavior and the effective unit of personality; it is generally agreed that people can be described by a limited number of traits. Although the range and specificity of traits are unique to each person, the conceptualization of "traits" is consistent and reflects regular individual behavior and features [13]. This final survey explored personality within five dimensions: extroversion, amenability, conscientiousness, neuroticism, and openness. The score for each dimension ranged between 5 and 30, with higher scores indicating that a person was more likely to be defined by that feature.

2.2 Grey Correlation Analysis Method

The grey correlation analysis is a quantitative description method that examines trends in the development of and changes in a system. The results reveal whether a connection is close by comparing the similarity between the reference data column and the geometry of the data column, which reflects the degree of correlation between curves.

Steps for grey correlation analysis:

- (1) Determine the analysis sequence

The reference sequence reflects the behavioral characteristics of a system, and is used to determine the comparison sequences that affect the behavior of that system. The data sequence that reflects the behavior of the system is known as the reference sequence, whereas the data sequence that affects the behavior of the system is called the comparison sequence.

Assuming that the reference sequence is $Y = \{Y(k) | k = 1, 2, \dots, n\}$; The comparative sequence can be assumed to be $X_i = \{X_i(k) | k = 1, 2, \dots, n\}, i = 1, 2, \dots, m$.

- (2) Determine the nondimensional variables

Because the data dimension of each factor is different, it is difficult to obtain a valid conclusion. Therefore, in our analysis of grey correlation degree, it was necessary to perform a dimensionless processing of the data. At present, the common nondimensional processing methods are extreme value, standardized, mean, and standard deviation, with the standardized method being the most common. However, if the indicators' means are all 0 and the standard deviations are all 1 when the indicators are obtained through standardized method, the results can only demonstrate the interaction between the indicators; by contrast, dimensionlessness equalizes the degrees of variation of the indicators, therefore, the standardization method does not apply to the comprehensive evaluation of multiple indicators. The covariance matrix of the data processed by the mean method can reflect the variance of each index in the original data, as well as information regarding the differences between each degree of influence of each index. Therefore, the mean method was adopted in this study.

$$x_i(k) = \frac{X_i(k)}{X_i(l)}, k = 1, 2, \dots, n; i = 1, 2, \dots, m$$

$X_i(l)$ represents the mean of the column i .

- (3) Calculate the correlation coefficient

The correlation coefficient between $x_o(k)$ and $x_i(k)$

$$\delta_i(k) = \frac{\min_i \min_k |y(k) - x_i(k)| + \rho \max_i \max_k |y(k) - x_i(k)|}{|y(k) - x_i(k)| + \rho \max_i \max_k |y(k) - x_i(k)|}$$

$\rho \in (0, \infty)$, is called the resolution coefficient. We usually take $\rho = 0.5$.

(4) Calculate the correlation degree

Because the correlation coefficient is the value of the correlation degree between the comparison sequence and the reference sequence at each moment (i.e., each point on the curve), the correlation coefficient has more than one value (it has a distinct value at each point on the curve); thus the information is too scattered to facilitate an overall comparison. Therefore, it is necessary to collect the correlation degree at each moment as a numerical value, and determine the mean value as a value of the correlation degree between the comparison and reference sequences. The correlation formula is as follows:

$$r_i = \frac{1}{n} \sum_{k=1}^n \delta_i(k), k = 1, 2, \dots, n$$

(5) Rank correlation degrees

The correlation degree is sorted by size, and if $r_1 < r_2$, the reference sequence y is more similar to the comparison sequence x_2 .

After calculating the correlation coefficient between $X_i(k)$ sequence and $Y(k)$ sequence, calculated the mean value of each kind of correlation coefficient. The average value r_i is called the correlation degree between $Y(k)$ and $X_i(k)$.

3 Analysis of the Results

Questionnaires were distributed to China Postal Airlines crew members, and were completed by people in various crew positions, including flight inspectors, teachers, pilots, and copilots. A total of 100 questionnaires were distributed; both the recycling and recovery rates were 50%, the overall efficiency was 100%.

3.1 Results of the Cargo Pilots Fatigue Survey

(1) Personal Information Survey

Of the cargo pilots who participated in this study, the age range was from 23 to 55 years; the average age was 32 years, the average work history was 8 years, and the average number of flight hours logged was 4385 h. Moreover, the percentage of the pilots who flew international routes, local routes, trunk routes, and feeder routes, respectively, was 23.4%, 29.9%, 36.4%, and 10.4%. In total, 53.1% of the pilots noted that their diet habits were regular, 49.0% did not drink, 61.7% did not smoke, and 87.8% identified a weekly exercise routine. Finally, the percentage of the pilots who believed that their health was in good, general, and poor condition, respectively, was 28.6% and 14.3%.

(2) Perceived Fatigue Survey

Perceived fatigue was divided into five levels: none, mild, moderate, severe, and extreme. According to the total scores, the following percentages indicate the

distribution of fatigue among the pilots: 8% perceived no fatigue, 23% perceived mild fatigue, 47% perceived moderate fatigue, 20% perceived severe fatigue, and 2% perceived extreme fatigue.

Fatigue was then divided into our set of dimensions, comprising general fatigue, physical fatigue, mental fatigue, and diminished motivation. The statistical distribution of the scores between the two sets of dimensions is presented in Table 1.

Table 1. Distribution of fatigue on different dimensions

Degree four dimensions	Free	Mild	Moderate	Severe	Extreme
General fatigue	2%	13%	33%	42%	9%
Physical fatigue	4%	14%	49%	22%	10%
Mental fatigue	4%	14%	43%	31%	8%
Reduced motivation	0%	16%	65%	14%	4%

After adding the second set of dimensions, it was determined that 51% of cargo pilots considered themselves to be severely fatigued, 32% considered themselves to be seriously physically fatigued, 39% considered themselves to be seriously mentally fatigued, and 18% considered their work motivation to be low.

A comparative investigation of the fatigue conditions of commercial pilots in China revealed that 6% of commercial pilots considered themselves to be severely and extremely fatigued, 16% considered themselves to be severely fatigued, 12% considered themselves to be physically fatigued, 18% considered themselves to be mentally fatigued, and 15% considered their work motivation to be very low. Therefore, in contrast to the commercial pilots, cargo pilots experienced more serious general, physical, and mental fatigue, and had lower work motivation.

(3) Work-Related Factors Survey

On the dimension of scheduling factors, we found that 55% of staff involved in flight investigations have adapted to their currently scheduled shifts and 45% have not adapted. Additionally, more than 90% of the pilots who identified as severely and extremely fatigued noted that their sleep habits were not compatible with their currently scheduled shifts. By contrast, 55% of the pilots overall felt fully rested after a night shift, of whom 77% were accustomed to their current shifts, whereas 45% of the pilots were unable to get adequate rest, of whom 78% were not accustomed to their current shifts. This analysis suggests that inadequate rest is a primary factor preventing pilots from adapting to their scheduled shifts; pilots who cannot adequately rest tend to perceive their workload as heavy.

On the dimension of workload factors, the statistical results showed that 55% of cargo pilots believed that they had heavy workloads. In total, 88% of the cargo pilots worked 4.2 days a week on average, and 61% worked 20 days each month and flew for more than 6 h over three segments per day. Moreover, 80% of cargo pilots suggested that night duty had a substantial influence on their fatigue, compared with only 4% of cargo pilots who believed that night duty had no effect or a minimal effect on their fatigue; thus, night duty was a key factor affecting the cargo pilot fatigue. Among the

68% of cargo pilots who were on duty for 4 or 5 days a week around the clock (i.e., before 12:00 am and after 4:00 am), cross-day work was more common.

On the dimension of working environment, we noted that people who work in a noisy environment, which is characterized by continuous background murmuring, are easily fatigued. Although new employees in such an environment can initially feel uncomfortable, they eventually adapt to this noise over time; this auditory noise eventually overstimulates the brain when a person's auditory threshold has risen sufficiently, which causes people to move slowly, think unclearly, and can lead directly to sleep [14]. Our survey results were consistent with these ideas. For example, we found that a plane's cabin environment impacts fatigue based on the conditions of noise, vibration, odor, temperature, and lighting, at rates of 89.8%, 53.1%, 53.1%, 40.8%, and 32.7%, respectively. To reduce cabin-environment-related fatigue, we suggest first addressing the distractions that are produced by noise, vibration, and odor, to obtain a multiplier effect. Our survey also revealed that 63% of cargo pilots flying at night tended to experience a strong sense of loneliness or depression.

On the dimension of other factors, our survey demonstrated that seasons are an important factor affecting fatigue, with 71% of cargo pilots noting that their fatigue levels were impacted by the season. In particular, 75% of the pilots revealed that their fatigue worsened in the summer, because summer flights need to avoid thunderstorms and thus the pilots faced larger workloads.

(4) Sleep Quality Survey

Sleep quality is also a crucial indicator of fatigue, and a lack of sleep is one of the main reasons that flight crews experience fatigue. In particular, cargo pilots who begin work early in the morning and work long hours experience increased degrees of fatigue. Notably, long-term sleep deprivation or poor sleep quality can lead to chronic fatigue, which not only impacts people's work but poses a marked threat to health. Table 2 shows the results of the sleep quality survey.

Table 2. Results of the sleep quality survey

Sleep conditions	The proportion
Sleep quality is very good	14%
Sleep quality is okay	50%
Sleep quality is mediocre	34%
Sleep quality is poor	2%

Specifically, the survey revealed that “mediocre” and “poor” sleep quality together accounted for 36%; only 14% of the pilots indicated that they had “very good” sleep quality.

Sleep quality was also assessed according to five dimensions of sleep status, namely sleep duration, sleep efficiency, sleep disorders, sleep medication use, and daytime dysfunction (see Tables 3, 4, 5, 6 and 7).

Table 3. Cargo pilots' sleep duration

The sleep time	The proportion
Less than 5 h	34%
5–6 h	12%
6–7 h	16%
More than 7 h	34%

Table 4. Cargo pilots' sleep efficiency

The sleep efficiency	The proportion
65%	38%
65%–74%	10%
75%–84%	16%
85%	36%

Table 5. Cargo pilots' night sleep disorders

Sleep disorders	The proportion
No	8%
Minor	54%
Larger	34%
Very difficult	4%

Table 6. Cargo pilots' daytime dysfunction

Daytime dysfunction	The proportion
No	6%
Minor	34%
Larger	34%
Very difficult	26%

Table 7. Cargo pilots who use sleep medicine

Using sleep medicine	The proportion
No	96%
Average 1–2 nights per week	2%
An average of more than 3 nights a week	2%

The statistical results indicated that 62% of the cargo pilots had an average sleep duration of less than 7 h per night. The percentage of pilots who slept nocturnally without disorders was 8%, while 6% slept during the day. 38% of the cargo pilots had sleep quality percentages that were less than 65%. However, only 4% of the pilots used

drugs to facilitate sleep. Overall, the data show that cargo plane pilots generally had poor sleep quality, which at least partially explains their high degrees of fatigue.

(5) Life Event Factors Survey

The life events analysis method provides a life cycle calculation period that covers 18 months, and accumulates the values of life changes that correspond to notable events that occurred during the cycle. The calculation of the total value of life changes is used as a statistical index; the higher the score is, the greater the degree of influence an event posed. The life events impact scores of the cargo pilots involved in this study are listed in Table 8.

Table 8. Scores from the life event factors survey

Scores	The proportion
0	50.0%
0 < score ≤ 50	46.0%
score > 50	4.0%

Notably, life events had no effect on 50% of the cargo pilots, a light impact on 46% of the pilots, and a substantial impact 4% of the pilots. Thus, the statistical data suggest that life events impact fatigue only to a limited extent.

(6) Personality Characteristics Survey

Different personalities tend to mitigate the stresses of life and work by different methods. Pilots’ sensitivity levels to the impact of fatigue may also reflect their personality characteristics. The statistical results of the personality characteristics survey adopted in this study are presented in Table 9.

Table 9. Results of the personality characteristics survey

personality characteristics \ Score	6-14	15-22	23-30
Extroversion	15.6%	80.0%	4.4%
accommodating	4.3%	78.3%	17.4%
conscientiousness	0.0%	57.8%	42.2%
neuroticism	31.1%	68.9%	0.0%
openness	2.2%	75.6%	22.2%

Specifically, we determined that the cargo pilots of freighter planes are more conscientious and open. Moreover, none of the cargo pilots had neurotic characteristics, indicating that they were interpersonally engaged, self-controlled, and emotionally stable.

3.2 Factors Affecting Cargo Pilots' Fatigue

According to the survey results, we determined that the following factors most crucially influenced cargo pilots' fatigue: age, total flight hours logged, total number of years as a pilot, diet, alcohol consumption, smoking, exercise routine, overall health status, sleep quality, adaptability to varying work shifts, workload self-assessment, number of working days per week, engagement in overnight flights, number of weekly flight hours, and the number of days of flight that spanned a day and night. The correlation coefficients of each variable, along with the degree of fatigue, were calculated using grey relational analysis. The results are listed in Table 10.

Table 10. Ranking of factors that affect fatigue

Influencing factors	Correlation degree	Ranking
Night flight	0.911	1
Workload	0.909	2
Health status	0.905	3
Number of working days per week	0.903	4
Whether sleep habits adapt to shifts	0.895	5
Exercise	0.892	6
Weekly flight hours	0.889	7
Age	0.878	8
The number of days of flight across the day and night	0.872	9
Regular diet or not	0.862	10
Sleep quality	0.860	11
Alcohol consumption	0.860	12
Smoking	0.856	13
Total flight years	0.794	14
Total flight hours	0.756	15

Subsequently, the results revealed that the factors associated with fatigue of airline pilots, in decreasing order, are: engagement in overnight flights, workload self-assessment, overall health status, number of working days per week, adaptability to varying work shifts, exercise routine, number of weekly flight hours, age, the number of flight days that spanned a day and night, diet, sleep quality, alcohol consumption, smoking, total number of years as a pilot, and total flight hours logged.

3.3 Comparative Analysis of Factors Affecting Fatigue Between Cargo and Airline Pilots

Eight of the factors that affect the fatigue of cargo pilots were also found to affect airline pilots, namely age, overall health status, workload self-assessment, sleep quality, total flying hours logged, number of working days per week, number of weekly flight hours, and total number of years as a pilot. The correlation coefficient along with

Table 11. Factors affecting the fatigue of airline pilots

Influencing factors	Correlation degree	Ranking
Work load	0.914	1
Health status	0.886	2
Age	0.881	3
Sleep quality	0.877	4
Number of working days per week	0.859	5
Weekly flight hours	0.841	6
Total flight hours	0.649	7
Total flight years	0.635	8

degree of fatigue, were calculated using the grey correlation method, and the results are shown in Table 11.

The analytical results revealed that the factors associated with fatigue in airline pilots, in decreasing order, are workload self-assessment, overall health status, age, sleep quality, number of working days per week, number of weekly flight hours, total flying hours logged, and total number of years as a pilot.

As revealed in Table 12, the effect of workload self-assessment, age, and sleep quality factors on fatigue does not significantly differ between airline and cargo pilots.

Table 12. Correlation coefficient comparison

Influencing factors	Airline pilots fatigue correlation degree	Cargo pilots fatigue correlation degree
Work load	0.914	0.909
Health status	0.886	0.905
Age	0.881	0.878
Sleep quality	0.877	0.860
Number of working days per week	0.859	0.903
Weekly flight hours	0.841	0.889
Total flying hours	0.649	0.756
Total flight years	0.635	0.794

In short, our comparative analysis indicated that workload self-assessment is the primary factor affecting airline pilots’ fatigue, whereas overnight flights are the primary factor affecting cargo pilots’ fatigue. However, the correlational degree ranking of overnight flights was higher than the workload self-assessment overall, indicating that overnight flights were more likely to make pilots feel fatigued than their workload self-assessment. Additionally, the number of working days per week and number of weekly flight hours were shown to have a greater effect on cargo pilots’ fatigue than on airline pilots’ fatigue, which may be related to the distinct work cycles of cargo and airline pilots. Specifically, most airline pilots are on duty during the day, whereas cargo

Table 13. Ranking comparison of the fatigue-influencing factors

Affecting factors of airplane pilots fatigue	Correlation degree ranking	Affecting factors of cargo pilots fatigue
Work load	1	Night flight
Health status	2	Workload
Age	3	Health status
Sleep quality	4	Number of working days per week
Number of working days per week	5	Whether sleep habits adapt to shifts
Weekly flight hours	6	Exercise
Total flying hours	7	Weekly flight hours
Total flight years	8	Age

pilots are mostly on duty at night; therefore, the effect of overnight flights on pilot fatigue is essential problem that must be addressed (Table 13).

4 Conclusion

1. Compared with commercial airplane pilots, the overall degree of fatigue experienced by cargo pilots is higher; additionally, the degree of both mental and physical fatigue is more serious among cargo pilots, with 51% of the pilots in this study self-identifying as severely and extremely fatigued.
2. The main reason pilots are unable to adapt to their scheduled shifts is that they do not get enough rest, which results in the perception of a heavier workload.
3. Noise and vibrations in the cabin environment have the greatest impact on fatigue. Pilots are also more likely to feel fatigued in the summer, and to feel lonely or depressed when flying at night.
4. The more prominent personality characteristics among cargo pilots are conscientiousness and openness, affinity for interacting with others, strong self-control, emotional adaptability, and emotional stability.
5. Night flights, a heavy workload, and overall health status are the three most critical factors associated with cargo pilots' fatigue; this knowledge can be used to effectively manage the problem of pilot fatigue.
6. No significant difference in workload self-assessment, age, or sleep quality is shown by the correlation degrees of cargo pilots' and airline pilots' fatigue. Thus, these three fatigue factors are considered to have a similar effect on fatigue, regardless of the type of pilot.

References

1. Balpa: HALF OF PILOTS HAVE FALLEN ASLEEP WHILE FLYING [EB/OL]. <http://www.balpa.org/News-and-campaigns/News/HALF-OF-PILOTS-HAVE-FALLEN-ASLEEP-WHILE-FLYING.aspx,2013-09-26>
2. Rosekind, M., Gregory, K.B., Miller, D.L.: Sleep quantity and quality of augmented long haul flight crews in onboard crew rest facilities. *Sleep Res.* **26**(7), 26–41 (1997)
3. open.nat.gov.tw/OpenFront/report_download.aspx?sysId=C09904149
4. Caldwell, J.A.: Fatigue in aviation. *Travel Med. Infect. Dis.* **3**(2), 85–96 (2005)
5. Caldwell, J.A., Mallis, M.M., Caldwell, J.L., et al.: Fatigue countermeasures in aviation. *Aviat. Space Environ. Med.* **80**(1), 29–59 (2009)
6. World Civil Aviation Safety Analysis Report. Civil Aviation Safety Research Institute [EB/OL]. <http://wenku.baidu.com/view/c806a864783e0912a2162a2a.html,2010-01>
7. Han, W., Hu, W., Wen, Z., et al.: Physiological and psychological factors and countermeasures of flight fatigue. *J. Fourth Mil. Med. Univ.* **29**(4), 379–381 (2008)
8. Ge, S., Wu, G., Xu, X., et al.: Effects of flight fatigue on visual fusion in civil aviation pilots of different ages. *Chin. J. Aerosp. Med.* **16**(3), 180–183 (2005)
9. Haiyan, L., Yongjian, L.: Analysis and countermeasure research of civil aviation flight fatigue. *Ind. Technol. Forum* **11**, 98–99 (2014)
10. Tianfang, W., Xiaolin, X.: Self-rating scale of fatigue. *Zhonghua Zhong Yi Yao Za Zhi* **24**(3), 348–349 (2009)
11. Smets, E.M.A., Garssen, B., Bonke, B.: etal The multidimensional fatigue inventory (MFI) sychometric qualities of an instrument to assess fatigue. *J. Psychosom. Res.* **39**(3), 315–325 (1995)
12. Buysse, D.J., Reynolds, C., Monk, T.H.: The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res.* **282**, 193–213 (1993)
13. Sun, R., Zhao, N., Li J.: Analysis of personality traits and behavioral safety of civil aviation pilots. *Ergonomics* **3**, 50–54 + 82 (2015)
14. Rui-shan, S.U.N., Wan-li, T.I.A.N.: Fly fatigue detection methods. *Occup. Health* **8**, 1142–1146 (2015)

A Landing Operation Performance Evaluation System Based on Flight Data

Lei Wang^(✉), Yong Ren, Hui Sun, and Chuanting Dong

Flight Technology College, Civil Aviation University of China,
Tianjin 300300, China
wanglei0564@hotmail.com

Abstract. Pilots' operation performance is closely correlated with flight safety, particularly in the final landing phase. The main purpose of this study is to develop a flight landing operation performance evaluation system based on flight data and a risk evaluation model. In this model, 3 flight parameters, including landing touchdown distance, vertical acceleration, and pitch angle of each flight, were used to objectively evaluate the performance of flight landing operations. The system is expected to be used to evaluate, analyze, and pre-alarm the performance of the landing operation of the pilot after the flight task, to provide practical technical support for airlines to monitor and control landing risk, and to provide a more accurate and objective basis for an airline's performance rewards and punishments.

Keywords: Landing safety · Flight operation · Flight data · Performance evaluation

1 Introduction

Pilots' operation performance can affect flight safety directly. Many studies have reported that pilot error is the primary cause of over 60% of flight accidents [1, 2]. The statistics on commercial flight accidents in China from 2006 to 2015 indicated that flight crew factors contributed to 64.58% of accidents [3]. Particularly in the final approach and landing stage, the occurrence rate of pilot error is significantly higher than in other phases because pilots need to deal with more situational change, greater decision making, and greater operational activity [4–6]. Accident statistics have also indicated that approach and landing was the most dangerous phase of flight; the landing phase in particular accounted for 23% of total fatal accidents occurring from 2006 to 2015, despite the fact that it accounts for just 1% of average flight time [7].

In the field of landing safety, many previous studies have focused on pilot visual perception and pattern analysis [8–12], runway overrun risk modeling [6, 13], critical factor analysis [14–16], and so on. Wang [17–19] applied a new method for landing safety research by using real flight data to analyze performance features of long landing incidents. However, there was relatively less research of landing performance. In particular, there were lesser outcomes regarding with performance and operation analysis based on real flight data.

The flight quick access recorder (QAR) is a system that can acquire aircraft operational data easily. It includes airborne equipment for recording data and a ground software station for storing and analyzing data. The QAR can record all kinds of aircraft parameters, pilot operation parameters, environmental features, and alarm information during an entire flight. The practice has proved that QAR data are helpful for improving flight safety management and quality control. However, the data have been rarely utilized in research.

The main purpose of this study is to develop a flight landing operation performance evaluation system based on flight quick access recorder data and a risk evaluation model. The system is expected to be used to evaluate, analyze, and pre-alarm the performance of the landing operation of the pilot after the flight task, to provide practical technical support for airlines to monitor pilots' landing operation performance and landing risk.

2 Methodology

2.1 Quick Access Recorder Data

QAR data can record the flight and operation of an airplane, information about its environment, and other types of information. The QAR data sampling frequency can reach as high as 16 Hz in modern aircraft. The Civil Aviation Administration of China (CAAC) has implemented the Flight Operations Quality Assurance (FOQA) program since 1997, with all commercial airplanes of Chinese airlines obliged to install a QAR or similar equipment [20]. Based on related operational rules and regulations, commercial airlines always use flight QAR data to monitor and analyze the entire aircraft and pilot operation performance in flight. When a flight parameter exceeds the prescriptive normal range, it is called a QAR Exceedance Event or Unsafe Event. Exceedance events usually do not lead to severe results, but they can increase the probability of an accident and bring potential harm to aircraft and even passengers.

Most flight operation departments in airlines utilize the flight data in a simple way. They generally just use the data to monitor flight safety status by counting the numbers of unsafe events. Obviously, this could not make full use of data resources through the simple logic of 'exceedance management'. Especially when the flight parameter of unsafe event is close to but not exceeds the threshold value, a large amount of data would be wasted. Therefore, a more effective method of using flight data is expected to be developed from the user-centered perspective.

2.2 Evaluation Model

There are large differences in the actual landing process. Pilots often must implement some emergency operations due to special circumstances, different pilots have different flight operation habits, and different aircraft also have different performance characteristics, so it is difficult to evaluate a pilot's operation performance directly. However, no matter the environmental factors, aircraft factors, and pilots' individual differences, the goal of the landing is the same, which is that the aircraft land in standard position

on the runway with the right pitch attitude, an appropriate speed, and an appropriate load. A pitch attitude exceeding the standard may result in tail strike incidents, a vertical load that is too heavy may result in a hard landing, and a pilot missing the standard landing point may cause an overrun runway accident. This means that pilots should try to reduce the risk of overrunning the runway, sustaining a hard landing, or causing tail striking when the aircraft touches down. Therefore, we introduce the concept of risk evaluation to evaluate the operation performance of pilots by calculating the risk of these 3 abnormal incidents after each landing. The risk value of these 3 landing incidents is taken as the index of landing operation performance evaluation.

Risk is a 2-dimensional concept that is usually measured based on 2 indicators. One is the severity of the consequences of an incident, and the other is the occurrence probability of an incident. The 3 parameters of touchdown distance, vertical acceleration, and pitch angle, which can be recorded by a QAR, are generally used as evaluation indexes for judging landing incidents. Theoretically, each kind of flight parameter distribution will be approximately a normal distribution in a period. If the distribution functions of the 3 parameters are obtained, it is possible to evaluate the severity of the consequences of landing incidents and the probability of these incidents occurring according to the distribution functions and the algorithms, and then the risk value of landing incidents can be calculated. Then, the final landing operation performance values can be calculated and combined with the weights of the 3 types of incidents. Based on the above analysis, the evaluation model of landing operation performance is written as follows:

$$\begin{cases} P_{landing} = \omega_1 \cdot R_{TD} + \omega_2 \cdot R_{VA} + \omega_3 \cdot R_{PA} \\ \omega_1 + \omega_2 + \omega_3 = 1 \end{cases} \quad (1)$$

$P_{landing}$ is the value of landing operation performance; ω_1 , ω_2 , and ω_3 are the respective weights of the touchdown distance, vertical acceleration, and pitch angle, which can be determined from statistical data of landing incidents or the Delphi method, or determined and adjusted by the safety supervision department of an airline according to the safety situation and monitoring strategy. R_{TD} , R_{VA} , R_{PA} are respective risk values of the 3 landing incidents (runway overrunning, hard landing, and tail striking). They can be calculated by formula 2 below:

$$\begin{cases} R_{TD} = P_{TD} \times S_{TD} \\ R_{VA} = P_{VA} \times S_{VA} \\ R_{PA} = P_{PA} \times S_{PA} \end{cases} \quad (2)$$

In formula 2, P_{TD} , P_{VA} , and P_{PA} represent the probability of the 3 landing incidents of runway overrunning, hard landing, and tail striking, respectively, and S_{TD} , S_{VA} , and S_{PA} represent the severity of the consequences of these respective landing incidents. These six parameters could be calculated out from the flight data distribution function of touchdown distance, vertical acceleration, and pitch angle. Then we have the value of landing operation performance combined with the weights of the 3 incidents, and the level of landing operation performance can be calculated out. The algorithm has been mentioned in our previous study [21].

3 System Design

In the last section, the landing operation performance evaluation model was established based on flight QAR data and risk evaluation theory. Three landing operation performance evaluation indexes (touchdown distance, vertical acceleration, and pitch angle) of each flight could be used to objectively evaluate the flight landing operation performance according to the model and algorithm. In this section, the flight landing operation performance evaluation system (FLOPES) will be introduced.

3.1 System Hierarchy

The flight landing performance evaluation system was designed to including 7 modules: flight data processing, operational performance evaluation, unsafe event inquiry, flight operation instructions, user center, user guidelines, system administration.

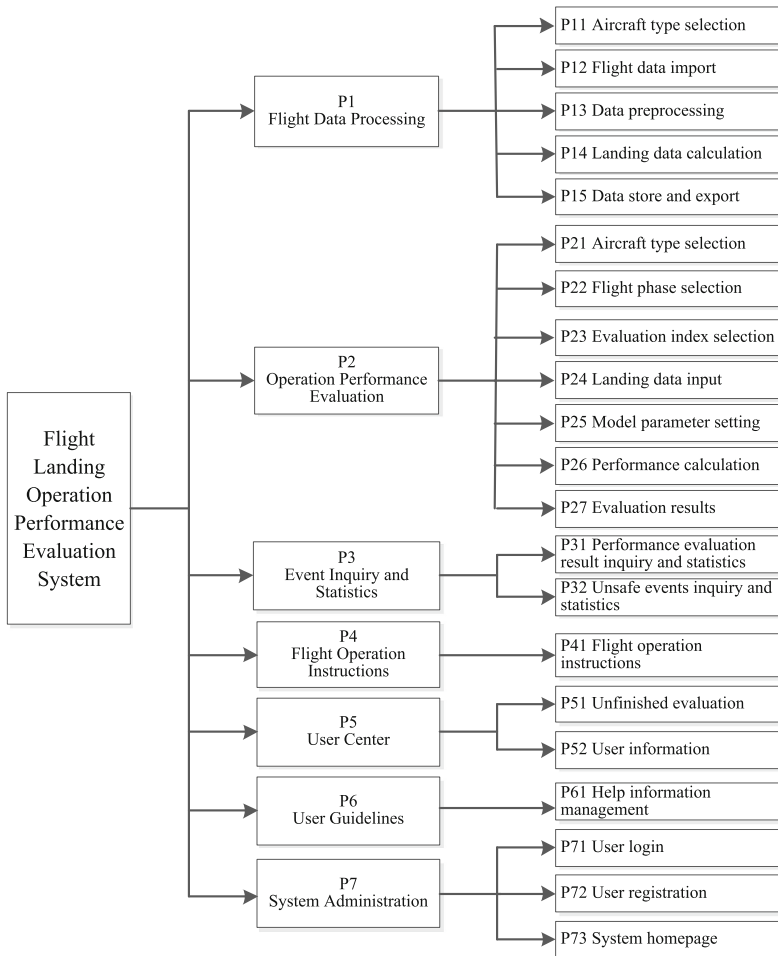


Fig. 1. Hierarchy diagram of flight landing performance evaluation system

and statistics, flight operation instructions, user center, user guidance, and system administration. The hierarchical structure of the system and each sub-function module of the system are shown in Fig. 1.

3.2 System Logic

The logic diagram of the flight landing performance evaluation system is shown in Fig. 2.

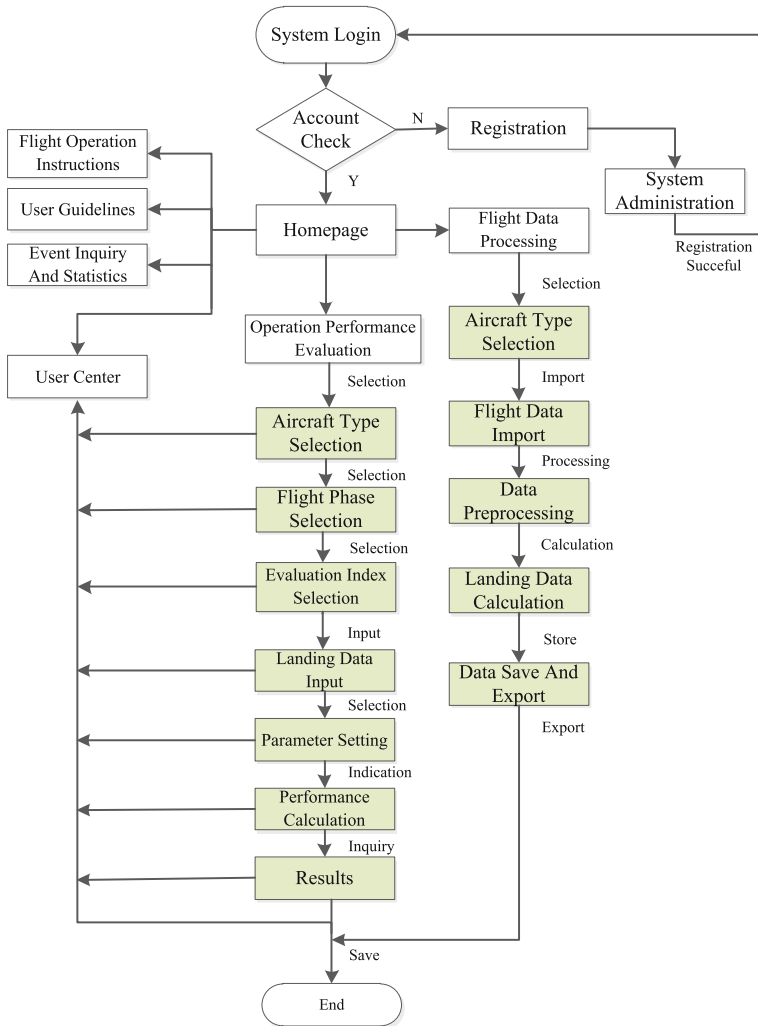


Fig. 2. Flow diagram of flight landing performance evaluation system

4 System Development

4.1 Development Environment and Process

Based on system design and database design, the MyEclipse tool is used to develop the flight landing operation performance evaluation system on the J2EE platform. Meanwhile, the SQL relational database management system is used to reduce network bottlenecks and improve data transmission efficiency.

4.2 System Interface and Functions

The developed Flight Landing Operation Performance Evaluation System (FLOPES) includes 7 modules, such as flight data processing, operational performance evaluation, and unsafe event inquiry and statistics. The main interface is shown in Fig. 3. The main interface includes a menu bar and links to 6 functional modules.

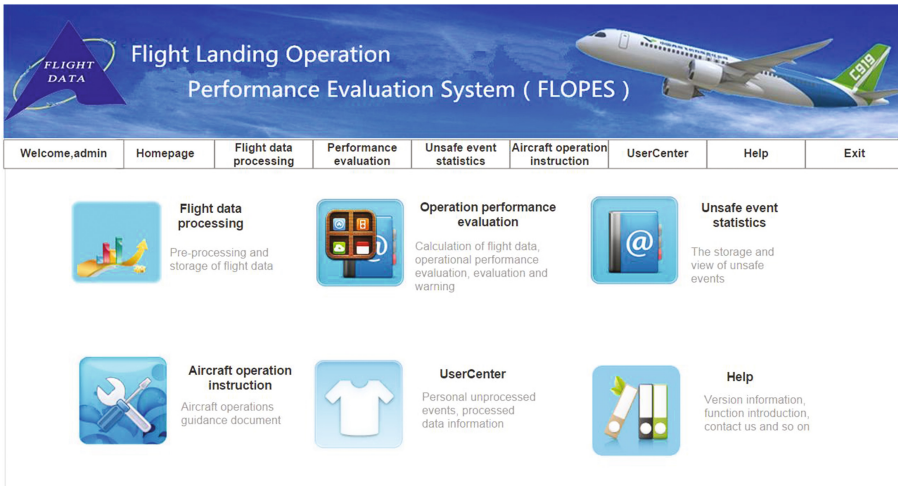


Fig. 3. Main interface of FLOPES

The core function of FLOPES is to evaluate landing operation performance using flight data. The entire evaluation algorithm of this function is illustrated in the last section. After clicking the “Operation Performance Evaluation” icon in the main interface and finishing the model parameters setting, the system will enter the calculation page. When the calculation is completed, the system will provide a prompt box and jump to the evaluation result page, as shown in Fig. 4. The evaluation results can be exported as Excel files for further use.

Another important function of FLOPES is to provide users with unsafe event inquiries and statistics. Users can enter the performance evaluation inquiry page and the

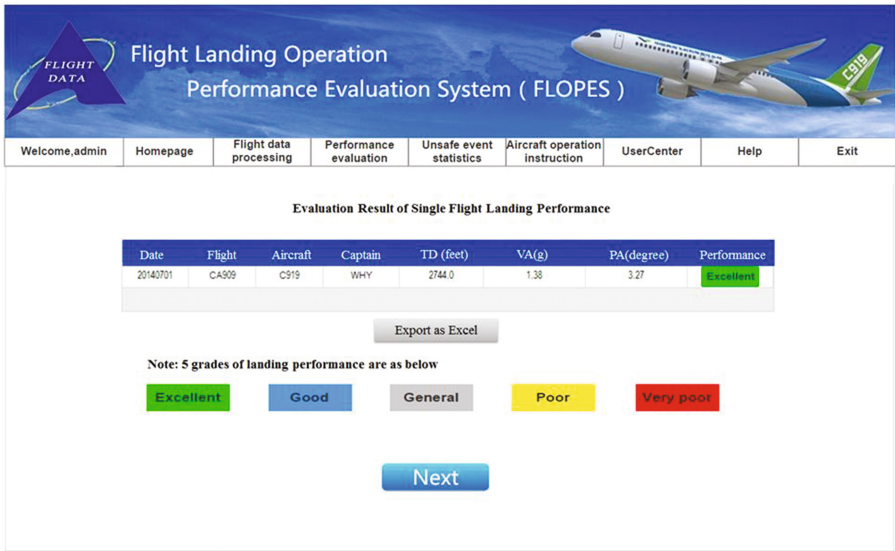


Fig. 4. Evaluation results of flight landing operation performance

Homepage -Unsafe Event Statistics

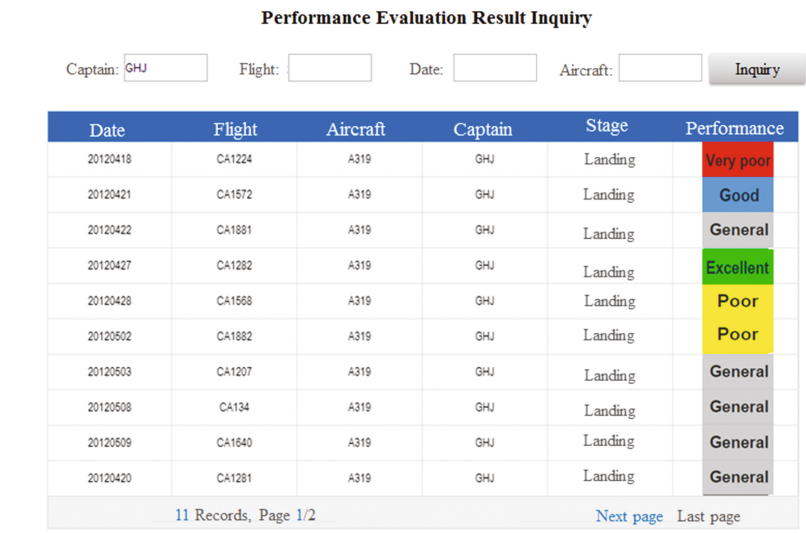


Fig. 5. Operational performance evaluation result inquiry and statistics

event inquiry statistics page by clicking “Unsafe Event Statistics” on the main interface. After inputting the information regarding the captain, flight date, flight number, and aircraft type, the system will indicate the relevant evaluation records and statistical results, as shown in Fig. 5.

Meanwhile, the system supports importation of QAR data into the system database and carrying out the corresponding inquiry. The user can inquire about relevant event information based on the entered information of captain, flight date, flight number, aircraft type, and event type or flight phase.

5 Conclusions

The Flight Landing Operation Performance Evaluation System was introduced in this study. The system was tested in the flight quality control department of an airline, and the QAR data of the flight was imported in batches to carry out flight data processing, landing operations performance evaluation, and other operations. The trial results showed the following:

- (1) The system can accomplish all basic functions, from the input of basic information and parameters to the output of evaluation results. It achieved landing operation performance evaluation, flight data processing, event inquiry and statistics, flight operation guidance management, user management, and other functions, indicating that the integrity of the system is good.
- (2) The system provides a support tool for flight operations quality assurance (FOQA) and flight training. The system can evaluate the performance of the landing operation of a flight that is more objective, effective, and reasonable than the simple overrun event management in current flight quality monitoring. It can provide a more accurate and objective basis for airline performance rewards and punishments.
- (3) The system provides actual data support for the flight operations department to monitor and control the landing risk. However, the system needs to be improved for shortening its response time when there is a mass data inputting and processing.

Acknowledgments. We appreciate the support of this work from the National Natural Science Foundation of China (No. 61304207), the Fundamental Research Funds for the Central Universities (No. 3122016B007) and the Graduate Technology Innovation Fund of Civil Aviation University of China (Y17-20).

References

1. Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., Wiegmann, D.: Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. *Hum. Factors* **49**(2), 227–242 (2007)
2. Jarvis, S., Harris, D.: Development of bespoke human factors taxonomy for gliding accident analysis and its revelations about highly inexperienced UK glider pilots. *Ergonomics* **53**(2), 294–303 (2010)
3. Civil Aviation Administration of China: Annual Report of China Aviation Safety. CAAC, Beijing, China (2016)

4. Wickens, C.D., Hollands, J.G.: *Engineering Psychology and Human Performance*, 3rd edn. Prentice Hall Press, Upper Saddle River (2000)
5. Stanton, N.A., Salmon, P., Harris, D., Marshall, A., Demagalski, J., Young, M.S., Dekker, S.: Predicting pilot error: testing a new methodology and a multi-methods and analysts approach. *Appl. Ergon.* **40**(3), 464–471 (2009)
6. Rosa, M.A.V., Fernando, G.C., Gordún, L.M., Nieto, F.J.S.: The development of probabilistic models to estimate accident risk (due to runway overrun and landing undershoot) applicable to the design and construction of runway safety areas. *Saf. Sci.* **49**(5), 633–650 (2011)
7. Boeing, Statistical summary of commercial jet airplane accidents, worldwide operations, 1959–2015. Boeing Commercial Airplanes, Seattle, WA (2016)
8. Galanis, G., Jennings, A., Beckett, P.: Runway width effects in the visual approach to landing. *Int. J. Aviat. Psychol.* **11**(3), 281–301 (2001)
9. Grosslight, J.H., Fletcher, H.J., Masterton, B., Hagen, R.: Monocular vision and landing performance in general aviation pilots: Cyclops revisited. *Hum. Factors* **20**(1), 27–33 (1978)
10. Wewerinke, P.H.: The effect of visual information on the manual approach and landing. National Aerospace Laboratory Technical Report (No. 8005U). NLR, Amsterdam, Netherlands (1980)
11. Palmisano, S., Favelle, S., Sachtler, W.L.: Effects of scenery, lighting, glideslope, and experience on timing the landing flare. *J. Exp. Psychol. Appl.* **14**(3), 236–246 (2008)
12. Jorg, O.E., Suzuki, S.: Modeling of the visual approach to landing using neural networks and fuzzy supervisory control. *Aerosp. Sci. Technol.* **14**(2), 118–125 (2010)
13. Kirland, I.D.L., Caves, R.E., Humphreys, I.M., Pitfield, D.E.: An improved methodology for assessing risk in aircraft operations at airports, applied to runway overruns. *Saf. Sci.* **42**(10), 891–905 (2004)
14. Reason, J.: *Human Error*. Cambridge University Press, New York, NY (1990)
15. Khatwa, R., Helmreich, R.L.: Analysis of critical factors during approach and landing in accidents and normal flight. *Flight Saf. Digest.* **17–18**, 1–256 (1999)
16. Li, W.C., Harris, D., Yu, C.S.: Routes to failure: Analysis of 41 civil aviation accidents from the Republic of China using the human factors analysis and classification system. *Accid. Anal. Prev.* **40**(2), 424–426 (2008)
17. Wang, L., Wu, C., Sun, R.: Pilot operating characteristics analysis of long landing based on flight QAR data. In: Harris, D. (ed.) EPCE 2013. LNCS, vol. 8020, pp. 157–166. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-39354-9_18](https://doi.org/10.1007/978-3-642-39354-9_18)
18. Wang, L., Wu, C., Sun, R.: An analysis of flight quick access recorder (QAR) data and its applications in preventing landing incidents. *Reliab. Eng. Syst. Saf.* **127**, 85–96 (2014)
19. Wang, L., Wu, C., Sun, R., Cui, Z.: An analysis of hard landing incidents based on flight QAR data. In: Harris, D. (ed.) EPCE 2014. LNCS, vol. 8532, pp. 398–406. Springer, Cham (2014). doi:[10.1007/978-3-319-07515-0_40](https://doi.org/10.1007/978-3-319-07515-0_40)
20. Civil Aviation Administration of China, Implementation and Management of Flight Operation Quality Assurance. Advisory Circular: 121/135-FS-2012-45. CAAC, Beijing, China (2012)
21. Wang, L., Wu, C., Sun, R., Cui, Z.: A quantitative evaluation model on hard landing risk based on flight QAR data. *China Saf. Sci. J.* **V24**(3), 1–10 (2014)

Dynamic Measurement of Pilot Situation Awareness

Xu Wu¹, Chuanyan Feng¹, Xiaoru Wanyan^{1(✉)}, Yu Tian²,
and Shoupeng Huang²

¹ School of Aeronautics and Engineering, Beihang University,
Beijing 100191, China

{wuxu0527, fengchuanyan, wanyanxiaoru}@buaa.edu.cn

² Astronaut Center of China, Beijing 100094, China

cctian@126.com, huangshoupeng2005@163.com

Abstract. This study mainly concentrated on ergonomics evaluation of the pilot situational awareness (SA) based on flight simulation platform and validation of prediction dynamic model according to the experimental result. The experiment scenario was designed as typical right-hand traffic pattern flight task. And situation awareness global assessment technique (SAGAT) method was used to measure the changing tendency of SA during the entire flight task, that the effectiveness of prediction model was verified by regression analysis. Moreover, online test of cognition capability was adopted to examine the relevance of SA under flight simulation task. The experiment revealed that the prediction model was validated with reasonable effectiveness, and SA of different subject varied, which was correlated with characteristics of cognition capability.

Keywords: Situational awareness · Dynamic measurement · Flight simulation task · Ergonomics experiment · Cognition capability

1 Introduction

The pilot SA was a vital index to evaluate the design of cockpit display interface. It was indicated that the pilot SA had a direct correlation to flight safety in relevant research. A correct judgment and decision could be made more rapidly and efficiently by pilot to achieve a higher level of flight safety when facing to higher SA [1, 2]. The statistical result of aviation accidents revealed that 35.1% non-major accidents and 51.6% major accidents were caused by the failure of pilot's decision-making, and the main reason was the lack of SA or SA error instead of the error in decision-making [3]. It was a common conception of SA in aviation ergonomics, but there still existed no strict definition. The three level definition presented by Endsley was a classic one and commonly accepted by other scholars [4].

It was considered that the operational definition of SA was mainly to measure the result of procedure (such as whether the event was comprehended by operator or not) instead of the process to the acquisition of relevant SA. For example, the pattern for pilot to aware the dangerous terrain was not an important issue, however, the

measurement of SA was only to evaluate whether it had been conscious of or not. So, the genuine measurement of SA should be concentrated on the dynamic element. The common measurements were situation awareness rating technique (SART), the measurement based on memory retention, the measurement based on operational performance, and the measurement based on physiological indices [5]. A combination of measurement based on memory retention, subjective assessment, and interview of critical event were employed by Paul to measure SA in experiment task [6]. The SAGAT method was used to measure SA by Riley, and meanwhile the immersion and mental workload was also investigated by questionnaire to analyze how the measurement of SA and attention allocation impacting the explanation of immersion cognition [7].

The conception of SA had already drawn extensive attention in all walks of life. The phenomenon not only related to the design of display supporting SA, but also to the happening reasons of disaster and accident. It should be ensured that the information which ought to be monitored in the design of interface should be presented to the pilots in a clear and comprehensible way under current automation. A tiny variation in the presenting format of information display would affect pilot SA. Therefore, the scheme of interface design should pay more attention to the factors with potential possibilities, and a full-scale experimental measurement would be made in multiple tasks [8]. The afterward measurement was the main method in current research, but the forecasting method was yet rarely that the prediction method based on the three levels of SA was even deficient. Moreover, the prediction and comprehend of pilot SA should be made to evaluate whether the interface design was good or bad in practical aviation industry, and this would be a scientific basis and theoretical foundation for further optimization of interface design and the reduction of human error.

2 Method

2.1 Subject

Nine graduated students from Beihang University were recruited as subjects in the experiment, with normal or corrected to normal eyesight, non-color blind and basic knowledge of civil aviation. They were all informed with the detail of experiment task and procedure, and voluntarily agreed to participate in the experiment.

2.2 Apparatus

The experiment platform was selected as simulation cockpit for flight simulation task, as shown in Fig. 1. The hardware of experiment platform was composed of main server, LED monitor, steering wheel, engine throttle, automatic control panel, seat and cockpit shell. The software of experiment platform was composed of flight simulation formula, which drove multi-screen to provide proximate actual experience of flight operation through simulation control devices.



Fig. 1. The experiment scenario

Table 1. Task operations of traffic pattern flight

No.	Transfer “Five-side” snapshot in control board and relieve freezing	Flight task phase	SAGAT question
1	Release braking, push the throttle until the engine N1 rotate speed was observed to 90%	Take-off and climbing phase	1–6
2	Observe airspeed on primary flight display, pull up at about 150 knots and then climb in 10~20 degrees pitching attitude		
3	Climb to about 1500 ft, gear up and retract the flaps		
4	Climb to about 2000 ft, disengage the steering wheel, connect the auto-throttle, airspeed hold, heading hold, and altitude hold switch, then turn on the autopilot		
5	Adjust heading hold switch from 179 to 269, turn to the second side		
6	Adjust heading hold switch from 269 to 359, turn to the third side, and keep on cruising	Cruising flight phase	7–19
7	Voice prompt, adjust heading hold switch from 359 to 89, and turn to the fourth side		
8	Turn on the radio navigation system, waiting for the automatic alignment to runway, and capturing glide slope		
9	Observe glide slope indicator (the pink rhombus to the third case), lighten APP mode, and then start to approach		
10	Flaps down slowly, altitude down to 2000 ft, flaps down to 25 degrees, and adjust airspeed to 180 knots	Approach and landing phase	20–23
11	Descend altitude to 1500 ft, adjust airspeed to 160 knots, gear down, and flaps down to 35 degrees		
12	Descend altitude to 1000 ft, adjust airspeed to 140 knots, and adjust airspeed to 130 knots		
13	When touched ground, click braking until shut down smoothly, close auto-throttle and auto-pilot, flaps up, and then finish all operations		

2.3 Experimental Design

Boeing 737–800 was chosen as experiment airplane, and weather was set as sunny day in summer. The flight task was selected as right-hand traffic pattern flight based on the current airport runway, which was initialized in Snapshot with magnetic heading 179, airspeed 220, auto-pilot heading hold 179, altitude hold 2800, and flap 5. The specific task operation was followed with the standard operation procedure according to flight manual of Boeing airplane, and was adjusted to the current situation of flight simulation platform.

The subjects were demanded to take several adaptive exercises of flight task till they were capable to complete the entire experiment task independently. SAGAT method was used to measure SA level in memory probing fashion, which required the subjects to make immediate response to the shown question as the flight task was randomly frozen. According to the purpose of current study, the experiment task was divided into take-off and climbing, cruising flight, as well as approach and landing. The relevant operation process was shown in Table 1.

2.4 SAGAT Measurement

SAGAT method was commonly used evaluation based on memory probing technology, which was applied to lead task process to randomly pause and substitute the display interface into questions, basically in choice form, that required the subjects to make immediate response according to the current status of task scenario. The specific question designed in the experiment was included with twenty-three task-related questions triggered by different conditions.

3 Results

Each question of SAGAT method in the experiment was referred to each key step of flight operations, involving with flight information on primary flight display, navigation display, and mode control panel. However, the content of SAGAT questions could be repeated, yet their triggering conditions were varied. And the experiment results were analyzed respectively by experiment task operations, and the tendency of accuracy and response time of SAGAT measurement was shown in Fig. 2.

In addition, the SA measurement was investigated separately according to each flight task phases mentioned in Table 1. For take-off and climbing phase, the dynamic results of SAGAT measurement were shown in Table 2.

Based on experiment results of accuracy and response time, the experiment value of SA was determined by accuracy and modified in consideration of response time. The specific quantification was shown below.

$$SA_i = \text{SAGAT accuracy}_i \times \frac{\text{SAGAT response time}_i}{\text{SAGAT average response time}} \quad (1)$$

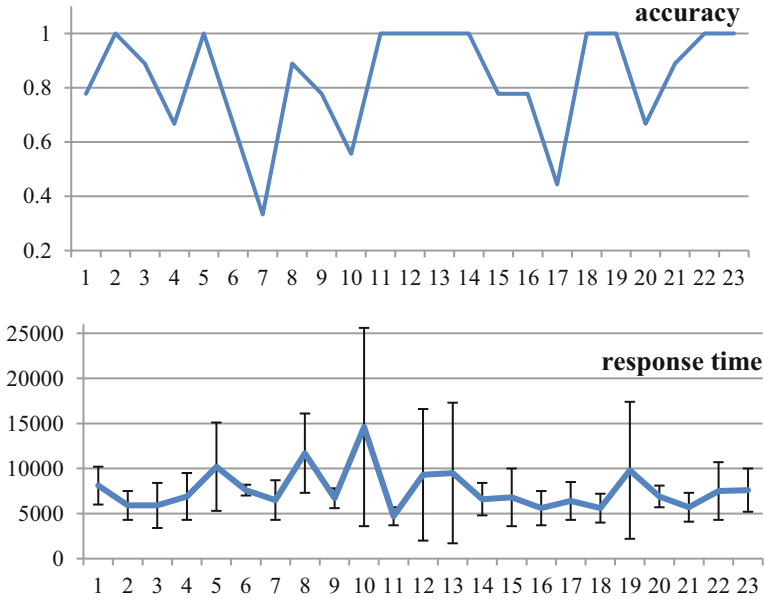


Fig. 2. The experimental results of accuracy and response time (ms) of SAGAT

Table 2. Results of SAGAT measurement during take-off and climbing phase

Indices	Push throttle	VR pull rod	Keep nose-up	Gear up and retract flaps	Engage autopilot
Accuracy	0.778	1.00	0.889	0.667	0.835
Response time(s)	8.1	5.9	5.9	6.9	8.9
Human error	0	0	0	2	1

And the experiment value of SA was calculated as 0.74, 1.30, 1.15, 0.75 and 0.72 according to Eq. 1. However, the situation might occur that such SA could be over 1.0, where the absolute value of SA would not be discussed because the purpose of this paper concentrated on the dynamic measurement of SA and the validation of prediction model. Therefore, the experiment value was proportionally normalized into 0 to 1.0, and that of take-off and climbing phase was shown in Fig. 3.

For cruising flight phase, the dynamic results of SAGAT measurement were shown in Table 3, and the modified experiment value of SA was shown in Fig. 4.

For approach and landing phase, the dynamic results of SAGAT measurement was shown in Table 4, and the modified experiment value of SA was shown in Fig. 5.

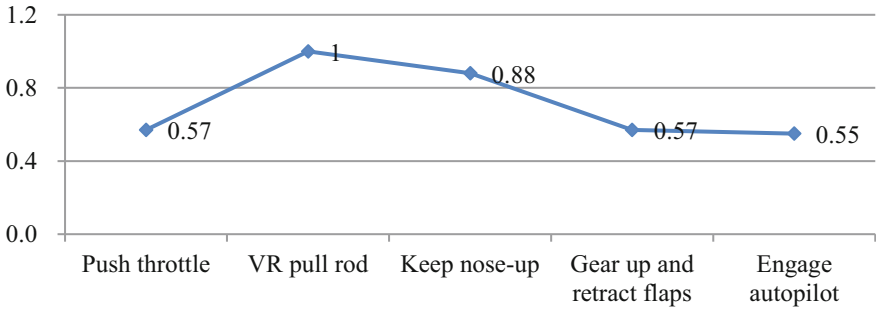


Fig. 3. Dynamic tendency of SA results during take-off and climbing phase

Table 3. Results of SAGAT measurement during cruising flight phase

Indices	Turn to the second side	Turn to the third side	The third side	Turn to the fourth side	Prepare to approach
Accuracy	0.610	0.640	1.00	0.780	0.853
Response time(s)	9.1	10.6	7.5	6.2	7.3
Human error	0	0	0	1	3

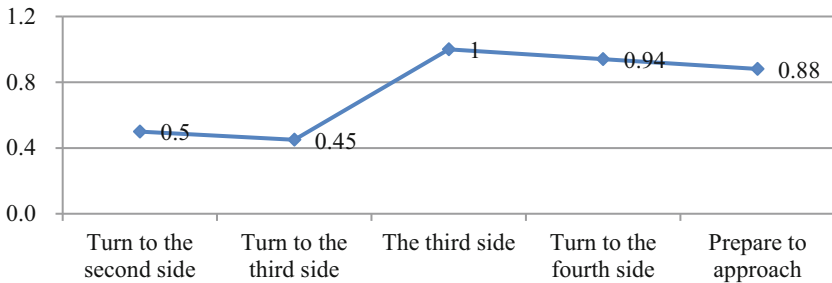


Fig. 4. Dynamic tendency of SA results during cruising flight phase

Table 4. Results of SAGAT measurement during approach and landing phase

Indices	Descend to 2000 ft	Descend to 1500 ft	Descend to 1000 ft	Landing
Accuracy	0.670	0.890	1.00	1.00
Response time(s)	6.9	5.7	7.5	7.6
Human error	1	0	0	0

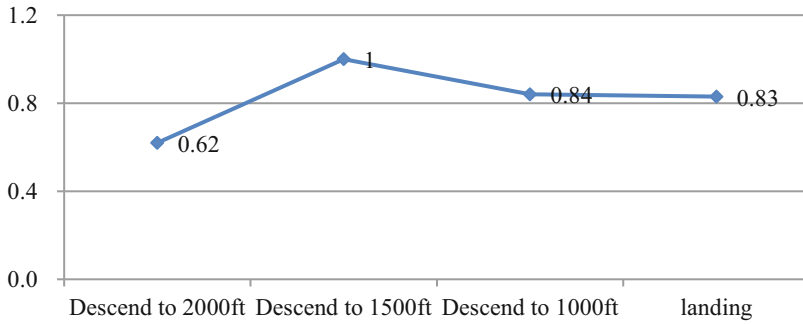


Fig. 5. Dynamic tendency of SA results during approach and landing phase

4 Validation of Prediction Model

In consideration of memory retain and attention allocation [9–11], the dynamic prediction model of SA was established by furthering Wu’s model [12], which illustrated the average level of SA influenced by each situational element at time T.

$$SA(t_i) = \sum_{i=1}^n (1 - 0.5k_i)u_iA_i \quad (22)$$

Where n was the number of situational element, u_i was the fuzzy membership of information priority, A_i was the attention allocated on situational element i, and K_i was individual difference concerned with the capability of information perception achieving to understanding.

According to experimental task design, the cockpit display interface was divided into nine situational elements: 1. airspeed, 2. altitude, 3. attitude, 4. heading, 5. navigation, 6. landing gear, 7. flap, 8. engine, 9. auto-pilot system. Taking “engage auto-pilot” (the last operation mentioned in Table 2) as example, the quantification of SA was shown below. For information priority, the situational element of No. 9 auto-pilot system was the highest, and priority ordering was followed with relevance between each situational element and current task operation: 9>3>5>2, 4>1>6, 7>8. For cognitive activation, the situational element of No. 9 auto-pilot system was also the highest as 1.0, and others was set as 0.5. And for memory retain, each situational element was declined with coefficient 0.9 or 0.8 based on its informativeness. The specific value of SA quantification was shown in Table 5, and the current SA was calculated as 0.473.

The dynamic prediction model was validated in regression analysis of experimental and predictable SA, respectively in different phases of take-off, cruising flight and landing. Both accuracy and response time of SAGAT measurement were concerned as experimental result of SA, and human error was also taken into account since it might be caused by misunderstanding of situational element. Therefore, the regression analysis of take-off and climbing phase showed that the prediction model achieved reasonable agreement with the experimental result with $R^2 = 0.863$, as shown in Fig. 6.

Table 5. Calculation of SA in each situational element

SA calculation	1	2	3	4	5	6	7	8	9
Information priority	0.08	0.10	0.14	0.10	0.12	0.06	0.06	0.04	0.30
Memory retention	0.512	0.205	0.64	0.205	0.329	0.9	0.9	0.656	1

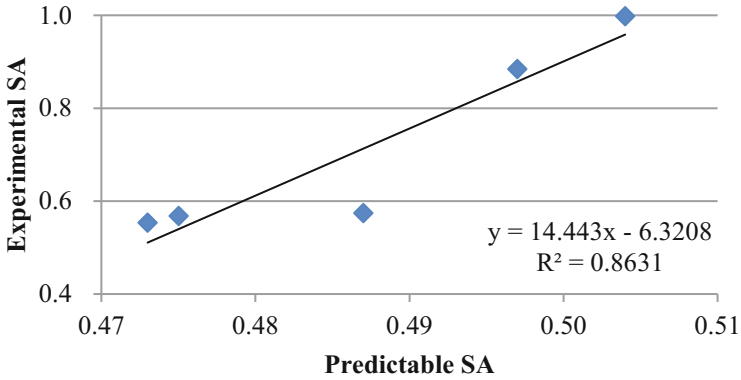


Fig. 6. Regression analysis of SA during take-off and climbing phase

And the regression analysis of cruising flight phase showed that the prediction model achieved certain agreement with the experimental result with $R^2 = 0.548$, as shown in Fig. 7. However, this experimental result was calculated without human error, it was modified by the occurring times of human error according to each task operations so that the experimental SA declined from 0.94, 0.88 to 0.88, and 0.70. Therefore, the regression analysis of the modified results showed better validation with $R^2 = 0.755$, as shown in Fig. 8.

Moreover, the regression analysis of approach and landing phase showed that the prediction model achieved considerable agreement with the experimental result with $R^2 = 0.835$, as shown in Fig. 9.

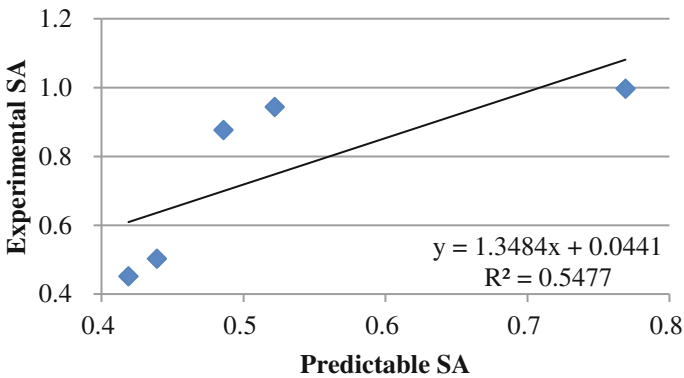


Fig. 7. Regression analysis of SA during cruising flight phase

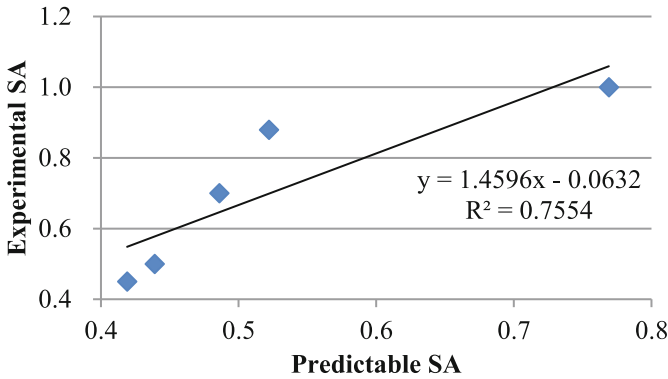


Fig. 8. Regression analysis of modified SA during cruising flight phase

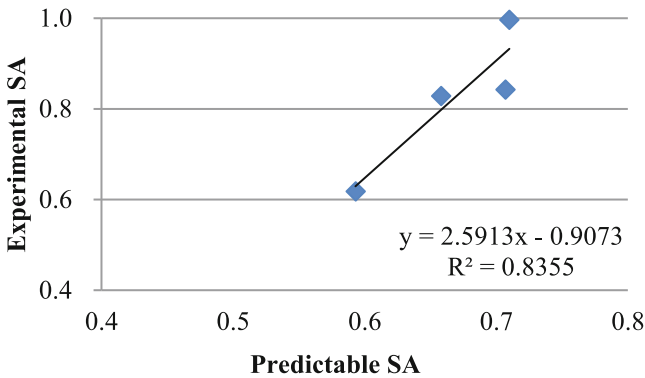


Fig. 9. Regression analysis of SA during approach and landing phase

5 Cognition Capability Testing

The influencing factors of SA included with both internal and external factors, and the internal ones were mainly related to the characteristics of perception and cognition capability as well as expertise skills. Since the subjects recruited in this experiment were equally trained and their experience of flight simulation task was almost the same, the influence of perception and cognition capability was primarily examined and analyzed, which was involved with basic response time, spatial rotation, short-term memory, and attention inhibition. The basic response time was tested in E-prime program according to classic paradigm of subtract method. And the others were tested online based on task of Rotation, Digit Span, and Double Trouble on the website of Cambridge Brain Sciences [13]. The subjects were required to take such tests accordingly after one entire exercise.

Table 6. Results of cognition capability testing

Simple reaction time (ms)	Selective reaction time (ms)	Discriminative reaction time (ms)	Spatial rotation (point)	Memory capability (point)	Attention inhibition (point)
267 ± 49	633 ± 117	451 ± 56	79 ± 23	10 ± 1	41 ± 15

Due to the small sample of nine subjects, the results of cognition capability were partially accorded with normal distribution except for short-term memory and simple reaction time, as shown in Table 6.

Correlation analysis was used to examine the relationship between cognition capability and experimental results of SA, which revealed significant correlation between response time of SAGAT measurement and simple reaction time ($r = 0.795$, $p = 0.010$), accuracy of SAGAT measurement and discriminative reaction time ($r = 0.702$, $p = 0.035$) as well as spatial rotation ($r = 0.704$, $p = 0.034$). Therefore, individual difference could be found between the subjects so that their SA varied in correlation with cognition capability.

6 Conclusion

In conclusion, the dynamic SA measured by SAGAT method showed good agreement with prediction model based on flight simulation task of right-hand traffic pattern flight. Moreover, the characteristics of cognition capability revealed significant correlation with SA.

Acknowledgement. This study was financially supported by Foundation of Key Laboratory of Science and Technology for National Defense (Program Grant No. 9140C770102140C77313).

References

1. Wei, H.Y., Zhuang, D.M., Wanyan, X.R., et al.: An experimental analysis of situation awareness for cockpit display interface evaluation based on flight simulation. *Chin. J. Aeronaut.* **26**, 884–889 (2013)
2. Endsley, M.R.: Situation awareness in aviation systems. In: *Handbook of Aviation Human Factors*, pp. 257–276 (1999)
3. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. *Hum. Factors* **37**, 32–64 (1995)
4. Endsley, M.R.: Errors in situation assessment: implications for system design. In: Elzer, P.F., Kluwe, R.H., Boussoffara, B. (eds) *Human Error and System Design and Management*. LNCIS, vol. 253, pp. 12–26. Springer, London, (2000)
5. Endsley, M.R.: Measurement of situation awareness in dynamic systems. *Hum. Factors* **37**, 65–84 (1995)
6. Salmon P., Stanton N., Walker G., et al: Situation awareness measurement: a review of applicability for C4I environments. *Appl. Ergon.* **37**, 225–238 (2006)

7. Jennifer, M.R., David, B.K., John, V.D.: Situation awareness and attention allocation measures for quantifying telepresence experiences in teleoperation. *Hum. factors Ergon. Manufact.* **14**, 51–67 (2004)
8. Endsley, M.R.: *Automation and Human Performance: Theory and Application*, pp. 163–181. Lawrence Erlbaum, Mahwah (1996)
9. Wickens, C.D., Jason, M.C., Thomas, L.: Attention–situation awareness (A-SA) model. In: *NASA Aviation Safety Program Conference on Human Performance Modeling of Approach and Landing with Augmented Displays*, pp. 189–205. NASA (2003)
10. Liu, S., Wanyan, X.R., Zhuang, D.M.: Modeling the situation awareness by the analysis of cognitive process. *J. Bio-Med. Mater. Eng.* **24**, 2311–2318 (2014)
11. Wu, X., Wanyan, X., Zhuang, D.: Pilot attention allocation modeling under multiple factors condition. In: Harris, D. (ed.) *EPCE 2013. LNCS*, vol. 8020, pp. 212–221. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-39354-9_24](https://doi.org/10.1007/978-3-642-39354-9_24)
12. Wu, X., Wanyan, X., Zhuang, D., Liu, S.: Pilot situational awareness modeling for cockpit interface evaluation. In: Harris, D. (ed.) *EPCE 2016. LNCS*, vol. 9736, pp. 476–484. Springer, Cham (2016). doi:[10.1007/978-3-319-40030-3_46](https://doi.org/10.1007/978-3-319-40030-3_46)
13. Cambridge Brain Sciences. <http://www.cambridgebrainsciences.com/>

An Approach for Assessing the Usability of Cockpit Display System

Hongjun Xue¹, Tao Li¹, and Xiaoyan Zhang^{1,2}(✉)

¹ Institute of Human Factors and Ergonomics,
Northwestern Polytechnical University, Xi'an 710072, China

xuehj@nwpu.edu.cn, 603649479@qq.com

² College of Mechatronics and Control Engineering,
Shenzhen University, Shen Zhen 518060, China

zxyliuyan@163.com

Abstract. Since the interaction between pilot and cockpit becoming more complicated and frequent, the usability of cockpit man-machine interface is directly related to the efficiency and safety of the cockpit. Among many human-computer interaction interfaces, the usability of the display system has become an important factor affecting flight efficiency and safety. However, the current study seldom pay sufficient attention on the usability evaluation. The limited research cannot report believable results to construct design. The paper is aimed to propose an evaluation model to evaluate the usability of displays.

To construct a quantitative model the first step is to establish a usability evaluation model composed of nine evaluation indicators. In this paper, factor analysis method is used to remove the overlapping factors. First of all, the raw data from the usability test was normalized to form a correlation matrix. Then the cumulative contribution rate of each factors is obtained from the eigenvalues of the matrix. And factors whose eigenvalues are greater than 1 are chosen as the primary index of the model. And then the paper establishes the factor load matrix, and the rotation load matrix is obtained by rotating it orthogonally. Removing the factors whose load less than 0.5, the rest of factors are chosen as the secondary index of the model. The multiple linear regression method is used to obtain the weight coefficient of every indicator in the usability evaluation model. The solution of equations is the weight coefficient matrix of each index. Score of the whole display system is figured out by weighting the score of each indicator. The evaluation result is the function of indicators of different displays built. Through the calculation and analysis of indicators, the usability of different systems can be acquired. Finally, the paper interchanges independent variables and the dependent variables, using linear regression analysis again. The validation and verification of the usability model had been executed by the questionnaires of flight simulation task based on typical flight scenes according to the A320 flight manual. The evaluation model of usability is helpful to the design of the new display system and the improvement of current system. And the evaluation model proposed in this paper can also be extended to evaluate the usability of other airborne systems and it will drive the development of civil aircraft cockpit usability.

Keywords: Cockpit display system · Usability evaluation · Factor analysis · Questionnaire

1 Introduction

The concept of usability emerged in the 1970 s, derived from the field of human-computer interaction. With the development of computer, the usability has been paid more and more attention from 1980 s. Usability is an important product quality attribute which was used to evaluate whether the product is easy to use and consist of user's needs and expectations [1]. Since then different definitions of usability have been provided, but the core is consistent that the user can use the product to meet certain needs [2]. User-centered design is the core methodology of usability engineering, which guides the design activities in all stages of the product development cycle. Usability design is a kind of user-friendly design. The designer always put the needs of users in mind and even invite the target user to participate in the design process. The usability design method can effectively improve performance and man-machine consistency, which greatly influence the success of the design. As an important measure of interaction design, usability describes the relationship between user and product. It is the evaluation method that users can use the product to achieve the final goal. The ultimate goal is user's ability to use the system well. That is to say, usability is a measure of product's quality from the user's perspective.

The cockpit is an extremely complex man-machine interaction system. In order to successfully complete the mission, the pilot must obtain a variety of flight attitude information through the display system first. Therefore, the cockpit display system usability level directly determines the flight safety and flight efficiency. However, modern aircraft cockpits expose the problem of mismatches and incongruities between aircraft design systems and people. Researchers have done a lot of research to improve the cockpit display system usability and the interaction between the pilot and the cockpit.

Fayollas [3] found a way to influence usability through reliability and apply it to the cockpit interface. A new generation cockpit, based on the ARINC 611 standard, allows the crew to control the display unit using a keyboard and cursor. Then, they analyzes the impact of system fault tolerance on usability and finds a trade-off between reliability and usability under complex training tasks.

Harbour [4] explored the relationship between situational awareness and display usability. Situational awareness is the ability to sense information and act in an acceptable way, and can directly affect the display of the HUD and the HDD, resulting in a change in the difficulty of the task. Harbour proposed the basic neurocognitive factors, determined its impact on the formation of SA, and studied the display usability. Visual attention, perceptual and spatial working memory were evaluated as predictors of different task difficulty, and the data of three predictors were statistically significant with the change of task difficulty.

Yanyan Wang [5] used the eye tracking technology to evaluate the usability of the static man-machine interface of the cockpit and established the static eye movement test data model. Based on the typical aircraft cockpit display interface, the participants need to find the corresponding position in the original image in the shortest possible time according to the given target image. The results show that there are 17 indicators of significant differences in the 25 commonly used eye movement index. After factor

analysis, five principal components were extracted, namely AOI, average pupil area, first glance distance and average fixed duration in AOI.

Yanbin Shi [6] proposed a cloud model to assess cockpit display system usability. The cloud model is composed of gray system theory and gray albedo function for qualitative and quantitative uncertainty transformation. They evaluated the human-computer interaction from the view of learning, efficiency, memory, error and satisfaction, and obtains the digital characteristics of each index in the cloud model. According to the model, it is possible to improve the usability of the flight simulator man-machine system by performing sensitivity analysis.

The study mentioned above has the following problems. Fayollas and Harbour explore the relationship between usability, reliability and situational awareness, but usability is not analyzed and evaluated separately. Wang Yanyan evaluates the various indicators using eye movements under static tasks, but in actual flight, dynamic display is more complex than static changes. Yanbin Shi's selection of evaluation indexes is very incomplete and has a great influence on the accuracy of the assessment results. Therefore, this paper is devoted to cover the shortages above. The paper will analyze the basic concepts of usability, and then select typical flight scenes, and evaluate the usability indexes of the cockpit display system comprehensively.

2 Method

2.1 Evaluation Process

There are many methods to study usability, but the most basic and useful is user testing, which has 3 basic components [7]:

1. Get hold of some representative users.
2. Ask the users to perform representative tasks with the design.
3. Observe what the users do, where they succeed, and where they have difficulties with the user interface.

In order to assess the usability of the cockpit display system, this paper select several graduate students as subjects. Refer to the A320 Flight Manual, takeoff, climb, descent, approach and landing is select as a typical mission. The remaining question is how to obtain the test data, select a more reasonable assessment factors and find the problem in the operation.

2.2 The Selection of Evaluation Indicators

There are a lot of factors influencing the usability of cockpit display system. The appropriate evaluation factors selected is key to evaluation accuracy and the result's effective for design. While more evaluation factors are selected, there will be increased overlap of indexes. However, if the factors are not sufficient, it would lead to the incomplete evaluation. These limitations would affect the accuracy of the evaluation and make the evaluation atypical. The challenge of the factors chosen is that it is always difficult to fully enumerate all the factors making the evaluation results not

comprehensive. In addition, it is difficult to construct a quantitative evaluation model directly, which will weaken the guidance to design.

Considering the usability indicators of cockpit display system are numerous and difficult to quantify, this paper try to use the questionnaire to collect test results. Comparing and screening the usability dimensions defined by the current standards such as ISO 9126, ISO 9241 and ISO 13407, indicators relating to cockpit display system usability are selected as shown in Table 1.

Table 1. Usability indicators of cockpit display system

Usability factors	Specific expression
Understandability	Easy to understand, correct syntax
Consistency	Font, color, layout of the display form and display location
Effectiveness	The completion of the task, useful
Efficiency	Fast and economic
Learnability	Easy to learn and remember
Errors	Tension, anxiety, unexpected situation
Recognizable	Clear identification
Satisfaction	Pleasure, satisfaction
Attractiveness	Appearance image, impression expression, aesthetic expression

At the same time, the following design factors are summarized according to the cockpit display system's own characteristics, as shown in Table 2. The design factor and the usability index are combined to construct the questionnaire for collecting the test results. Each small problem of the questionnaire is the evaluation factor.

Table 2. Design factors of cockpit display system

Category	Design factors
Appearance layout	overall shape, relative position, angle
Interface	interface elements, the instrument arrangement
Interface properties	color settings, resolution, viewing angle, brightness, icon size

2.3 Factor Analysis and Multiple Linear Regression

Factor analysis is used to combine and classify the factors which overlap each other, and obtain the factors that are not related to each other. The essence of factor analysis is to reduce the number of significant variables, the main mathematical thought is to reduce dimension and simplify, and the main purpose is to find a small number of essential factors. The calculation procedure of the factor analysis method is as follows:

1. Normalization of raw data: Make the data positive and non-dimensional
2. Determine the number of common factors: to obtain the correlation coefficient matrix and its eigenvalues and eigenvectors, the contribution rate of accumulation

3. Solving the factor load matrix: the meaning of each element in the factor load matrix represents the similarity coefficient between the corresponding variable and the common factor, reflecting the relative importance between them. The higher the absolute value is, the higher the correlation degree is.
4. Rotation of the factor load matrix: simplify the structure of the factor load matrix

After determining the final evaluation factors, the paper use multiple linear regression analysis to determine the weight coefficient. Multivariate linear regression analysis is a mathematical statistical method to deal with the statistical correlation of variables. Multivariate Regression Assume that the dependent variable is a multivariate linear function of the independent variable, and the stepwise regression method is used to screen these independent variables to find the statistical multiple linear regression model. The basic idea of multivariate regression analysis is that although there are no deterministic functional relationships between multiple independent variables and dependent variables, the mathematical expression that best reflects their relationship can be found. An important use of multiple linear regression analysis is to interpret and predict the data, and to calculate the accuracy of interpretation and prediction.

3 Usability Testing

3.1 Tasks

During the course of the experiment, the participants need to complete the whole process of the aircraft driving from takeoff to landing, and then fill out the questionnaire according to the subjective experience of flight test. The questionnaire included 35 declarative sentences, and the participants were evaluated according to nine levels from “very dissatisfied” to “very satisfactory”.

3.2 Participants

There are nine testers involved in the experiment and all of whom are graduate students who have been studying usability for one or two years. They are between 22 and 25 years old, with an average age of 22.8 years, so we can assume that they are appropriate participants. Although the participants have obvious deficiencies in the control and control ability compared with the pilot, they have the ability of cognition and judgment on the cockpit human-computer interaction interface. The satisfaction questionnaire was conducted under the guidance of the main test, which ensured the validity of the questionnaire.

3.3 Apparatus

The test is provided with a cockpit simulator, which is a simplified version, but the basic functions on the display are complete (Fig. 1).

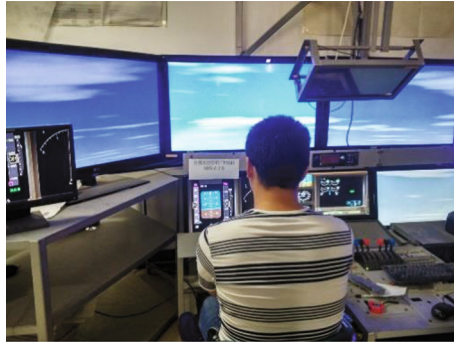


Fig. 1. The cockpit simulator

3.4 Results

As the amount of experimental data is large, the original survey data is no longer listed here. The scores of each index in the questionnaire were statistically analyzed, and the cumulative contribution rate was obtained by factor analysis using PASW in Table 3.

Table 3. The cumulative contribution rate of each factor

Ingredients	Initial eigenvalue		
	Sum	Variance%	Cumulative contribution%
1	4.270	47.444	47.444
2	2.052	22.797	70.241
3	1.471	16.342	86.584
4	0.690	7.672	94.255
5	0.261	2.897	97.152
6	0.191	2.122	99.273
7	0.046	0.515	99.788
8	0.019	0.212	100.000
9	3.286E-17	3.651E-16	100.000

As can be seen from the table, there are three main indexes affecting the usability of the cockpit display system, whose cumulative contribution rate reaches 86.584%. And the rotation component matrix is obtained in Table 4 after the rotation of the factor load matrix.

In the rotation component matrix, factors whose load is less than 0.5 are deleted. And the three main indexes that affect the usability of the cockpit display system are finally determined through summarizing the rest of the factors in the table. The final three main factors are satisfaction, learning and error.

Table 4. Rotation component matrix

Factors	Ingredients						
	1	2	3	4	5	6	7
VAR00002	0.905	0.187	-0.266	0.230	0.083	-0.002	0.038
VAR00033	0.877	-0.276	0.070	0.060	-0.018	0.018	0.375
VAR00032	0.733	0.437	-0.284	0.239	0.297	0.022	0.072
VAR00013	0.629	0.408	0.104	0.350	0.073	-0.489	-0.128
VAR00030	-0.131	0.919	0.083	-0.077	-0.076	0.124	0.104
VAR00015	0.238	0.904	0.239	0.164	0.150	-0.102	0.026
VAR00016	-0.308	0.831	-0.421	0.101	0.037	-0.108	-0.057
VAR00005	-0.049	0.624	0.460	0.010	0.434	0.208	0.200
VAR00018	-0.014	-0.207	0.935	0.137	0.034	0.166	0.157
VAR00026	0.407	-0.116	-0.862	-0.154	0.181	-0.082	-0.089
VAR00020	0.376	-0.270	-0.853	-0.146	-0.162	0.057	-0.080
VAR00012	0.497	0.414	0.575	0.049	-0.176	0.410	0.211
VAR00021	-0.093	-0.060	-0.135	0.924	-0.305	-0.111	0.025
VAR00025	-0.072	-0.079	0.349	0.921	-0.075	-0.028	0.099
VAR00007	0.216	0.298	-0.029	0.772	0.219	0.220	0.384
VAR00019	0.126	0.470	0.302	0.643	0.344	0.304	0.209
VAR00017	-0.451	-0.171	-0.079	-0.134	-0.601	-0.457	0.351
VAR00001	0.015	0.094	-0.030	0.250	0.096	0.952	0.087
VAR00008	0.392	0.172	-0.283	0.321	0.336	-0.686	-0.178
VAR00004	-0.291	-0.211	-0.211	0.319	-0.455	-0.675	-0.248
VAR00034	0.270	0.382	0.446	0.067	0.375	0.596	0.263
VAR00010	0.205	0.263	0.340	0.269	0.129	0.010	0.827
VAR00009	0.072	-0.021	0.128	0.170	-0.044	0.385	0.824

Table 5. The weight of the three factors

Factor	Learnability	Errors	Satisfaction
Weights	0.152	0.507	0.605

Table 6. Comparison table of design factors

Factors	Weight	Design factors
VAR00009	0.628	Icon size
VAR00017	0.539	The instrument arrangements
VAR00018	0.287	Interface element
VAR00005	0.274	Position
VAR00012	0.26	Color settings
VAR00015	0.165	
VAR00006	0.076	
VAR00020	-0.215	Brightness

Multivariate regression analysis can determine how cockpit display system usability is affected by factors in the model. Three main factors of the model are independent variables while cockpit display system usability is the dependent variable. The results of the regression analysis are shown in Table 5.

Take satisfaction for example, comparing its regression model with the evaluation factors can finally find design factors which influence satisfaction. The results are shown in Table 6.

4 Conclusion

The evaluation model proposed in this paper is verified. Three main factors influencing the usability of the cockpit display system are proposed, and the evaluation index system is established and a better evaluation result is obtained. Furthermore, this paper finds the design factors behind the indexes, and the problems in the cockpit display system design are found. Moreover, the evaluation model established in the paper can be applied to the whole cockpit usability assessment. This model can also be extended

to other man-machine interface usability assessment, and promote the development of display system interface usability.

With the development of cockpit display system technology, we will continue to improve the experimental program, adjust the evaluation factors and indicators to optimize the evaluation model to get a more accurate assessment results.

References

1. Nielsen, J.: Usability engineering. Elsevier, Amsterdam (1994)
2. Andon, C.L.: Usability analysis of wireless tablet computing in an academic emergency department: capstone project (2004)
3. Fayollas, C., Martinie, C., Palanque, P., et al.: An approach for assessing the impact of dependability on usability: application to interactive cockpits. In: Tenth European Dependable Computing Conference, pp. 198–209. IEEE Computer Society (2014)
4. Harbour, S.D., Christensen, C.A.: Neuroergonomic quasi-experiment: predictors of situation awareness and display usability while performing complex tasks. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series (2015)
5. Wang, Y., Liu, Q., Xiong, D., et al.: Research on assessment of eye movement sensitivity index through aircraft cockpit man-machine interface based on eye movement tracking technology. In: Proceedings of the 15th International Conference on Man-Machine-Environment System Engineering, pp. 495–502 (2015)
6. Shi, Y., Ouyang, D.: Usability Evaluation of the Flight Simulator's Human-Computer Interaction. In: Intelligent Computing Theories and Application (2016)
7. Shackel, B.: Usability-context, framework, definition, design and evaluation. *Interact. Comput.* **21**(5–6), 339–346 (2009)
8. Mayer-Ullmann, D., Weber-Schaefer, U., Held, T.: Testing usability of a software program. US, US 7673287 B2 (2010)
9. Wang, L., Cao, Q., Chang, J., et al.: The effect of touch-key size and shape on the usability of flight deck MCDU, pp. 234–238 (2015)
10. Grossman, T., Fitzmaurice, G., Attar, R.A.: Survey of software learnability: metrics, methodologies and guidelines. In: International Conference on Human Factors in Computing Systems, pp. 649–658 (2009)
11. Theuma, K., Cauchi, N., Gauci, J., et al.: Design and evaluation of a touchscreen concept for pilot interaction with avionic systems. In: Digital Avionics Systems Conference (2015)
12. Kieras, D., Polson, P.G.: An approach to the formal analysis of user complexity. *Int. J. Man Mach. Stud.* **22**(4), 365–394 (1985)
13. Sears, A., Jacko, J.A., Chu, J., et al.: The role of visual search in the design of effective soft keyboards. *Behav. Inform. Technol.* **20**(3), 159–166 (2001)

Cognition and Driving

Partial-autonomous Frenzy: Driving a Level-2 Vehicle on the Open Road

Francesco Biondi^(✉), Rachel Goethe, Joel Cooper, and David Strayer

University of Utah, Salt Lake City, UT, USA
francesco.biondi@utah.edu

Abstract. Partial-autonomous vehicles are among us and represent a prominent testing ground for assessing the human interaction with autonomous vehicles. One main limitation of the studies investigating *would-be* users' attitude toward partial to full autonomous driving stems from their indirect experience with such technology. In this study, participants drove a partial-autonomous vehicle on the open road and interacted with both Adaptive Cruise Control (ACC) and Lane Keeping Assist (LKAS) systems. Preliminary results show participants rating level-2 autonomous features as possible sources of stress. Participants had issues engaging these systems with denser traffic and thought these systems to be more beneficial in traffic-free driving. Compared to ACC, engaging LKAS and monitoring its functioning represented a more challenging task and participants' ratings of stress toward this system increased over time. Findings obtained in this study are of importance for exploring user interaction with future highly-autonomous vehicles and designing effective countermeasures to make the human-machine interface of these systems more informative and easier to use.

Keywords: Autonomous vehicles · Trust · Acceptance · Partially autonomous · Highly autonomous · Human-machine interface

1 Introduction

Partial-autonomous vehicles are among us and represent a prominent testing ground for assessing the human interaction with autonomous vehicles. The Society for Automotive Engineers [1] defines five levels of driving automation based upon the system's capability to execute lateral and longitudinal maneuvers, monitor the driving environment, respond to emergencies and drive without the aid of the human driver in various traffic scenarios. Whilst we expect vehicles with level-3 to level-5 autonomous capabilities to gradually hit the market in the next 2 to 25 years [2], level-2 vehicles – i.e., vehicles equipped with systems capable of executing steering and acceleration operations but requiring the human driver to monitor the traffic environment, are currently being driven on US roads.

Over the last fifteen years, the volume of studies investigating public's attitudes toward autonomous systems has grown exponentially. Topics such as trust and acceptance toward vehicles with autonomous capabilities have been the focus of

investigation in the driving community. In the study by Koo et al. [4] authors manipulated the type of information provided to the driver by a collision avoidance system and investigated its effect on trust. Whenever a possible collision was detected, the collision avoidance system automatically applied the brakes and one of four possible warning messages was presented to the driver. The four warning messages contained information regarding: the behavior of the system (i.e., “car is braking”; *How* message), the reason for the system to intervene (i.e., “Obstacle ahead”; *Why* message), a combination of the two (i.e., “Car is braking due to obstacle ahead.”; *How + Why* message), or none of the above (no message). Results showed that the content of the message had a significant effect on ratings with participants surprisingly feeling less positive about the system in the *How + Why* compared to the remaining three conditions. Interestingly, highest ratings of trust were found in the *Why* condition where the information provided to the drivers was more salient and, thus, quicker to process in the context of near collisions. More recently, Itoh et al., [3] investigated participants’ interaction with a simulated semi-autonomous vehicle equipped with an auto braking system. Results showed that participants positively evaluated the assistance system with respect to its value in avoiding collisions. In particular, feelings of acceptance increased in high emergency scenarios.

One possible limitation of the studies investigating *would-be* users’ attitude toward partial to full autonomous driving stems from their indirect experience with such technology. In many studies information regarding user interaction with partially to fully autonomous systems was collected either in simulated scenarios [5] or via surveys [6, 8]. In this study, we investigated user interaction with partially-autonomous vehicles and observed how direct exposure to level-2 vehicles shaped the user experience of such systems. Further, given the limited availability of highly and fully autonomous vehicles, testing a level-2 vehicle on the open road will allow to identify challenges characterizing the user interaction with partial automation and, thus, develop possible solutions for level-4 and 5 vehicles.

The aim of this study is to investigate how naïve drivers interact with level-2 vehicles and, in particular, assistance systems such as Adaptive Cruise Control and Lane Keeping Assist. In this study, participants drive a partial autonomous vehicle on highway roads for one hour and interact with ACC and LKAS using the human-machine interface (HMI) available on the 2016 Honda Accord equipped Honda Sensing. Subjective ratings of trust and acceptance as well as comments regarding the overall interaction with the systems are collected.

2 Method

2.1 Participants

Ten participants (one male) participated in this study. Participants had an average age of 25 years (standard deviation: 3 years). All participants had normal neurological functioning, normal or corrected-to-normal visual acuity, normal color vision, a valid

driver's license, and were fluent in English. Participants admitted they did not have any prior direct experience driving vehicles with ACC and LKAS. A University of Utah Institutional Review Board (IRB) authorization.

2.2 Materials

Participants drove a 2016 Honda Accord equipped with Honda Sensing. Honda Sensing is a suite of advanced driver assistance systems (ADAS) including: Collision Mitigation Braking System – an automatic braking system that applies the brakes whenever a collision is deemed unavoidable, Adaptive Cruise Control, Lane Keeping Assist System, Road Departure Mitigation – a system designed to help steer to help keep the vehicle from leaving the road, Lane Departure Warning – a system designed to monitor vehicle lane position and alert when it drifts into a new lane without the driver signaling, Forward Collision Warning, - a system integrated with Collision Mitigation that is designed to detect the presence of vehicles in front and alerts the driver when approaching at high speed (see Fig. 1).



Fig. 1. 2016 Honda accord XL with Honda sensing used in this study

SAE describes level-2 vehicles as following: “the driving mode-specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the human driver performs all remaining aspects of the dynamic driving task” (see Fig. 2). The 2016 Honda Accord equipped with Honda Sensing thus qualifies as SAE level-2 vehicle.

A GoPro camera was used during the drive for recording the dialogue inside the cabin. The questionnaire (Questionnaire 1) used to collect subjective ratings was composed by 10-point Likert scales measuring: overall opinion toward ACC and LKAS, trust, acceptance, usability, ease-of-use, stress, feeling of safety.

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/Deceleration	Monitoring of Driving Environment	Fallback Performance of Dynamic Driving Task	System Capability (Driving Modes)
Human driver monitors the driving environment						
0	No Automation	the full-time performance by the <i>human driver</i> of all aspects of the <i>dynamic driving task</i> , even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	Driver Assistance	the <i>driving mode</i> -specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the <i>human driver</i> perform all remaining aspects of the <i>dynamic driving task</i>	Human driver and system	Human driver	Human driver	Some driving modes
2	Partial Automation	the <i>driving mode</i> -specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the <i>human driver</i> perform all remaining aspects of the <i>dynamic driving task</i>	System	Human driver	Human driver	Some driving modes
Automated driving system ("system") monitors the driving environment						
3	Conditional Automation	the <i>driving mode</i> -specific performance by an <i>automated driving system</i> of all aspects of the <i>dynamic driving task</i> with the expectation that the <i>human driver</i> will respond appropriately to a <i>request to intervene</i>	System	System	Human driver	Some driving modes
4	High Automation	the <i>driving mode</i> -specific performance by an automated driving system of all aspects of the <i>dynamic driving task</i> , even if a <i>human driver</i> does not respond appropriately to a <i>request to intervene</i>	System	System	System	Some driving modes
5	Full Automation	the full-time performance by an <i>automated driving system</i> of all aspects of the <i>dynamic driving task</i> under all roadway and environmental conditions that can be managed by a <i>human driver</i>	System	System	System	All driving modes

Copyright © 2014 SAE International. The summary table may be freely copied and distributed provided SAE International and J3016 are acknowledged as the source and must be reproduced AS-IS.

Fig. 2. Levels of autonomy (Society for Automotive Engineers (2014))

2.3 Procedure and Design

Before the study began, participants filled out a consent form approved by the University of Utah Institutional Review Board. Once in the vehicle, participants provided information regarding their demographics and knowledge of autonomous vehicles. We adopted a pre-post experimental design with participants completing questionnaire 1 before and after driving the level-2 vehicle and experiencing ACC and LKAS. This allowed us to measure how their knowledge and attitude toward ACC and LKAS changed over time with experience of the systems. Before driving, participants completed questionnaire 1 for the first time (pre). After completing the questionnaire, a general overview of the commands to be used for controlling ACC and LKAS was provided to participants by the research assistant sit next to them. We did so to help participants familiarize with the new vehicle and HMI of ACC and LKAS. The starting point for the drive was in the parking lot of the Department of Psychology at the University of Utah. After 15-min drive in a residential area of Salt Lake City, Utah, participants entered the eastbound I-80 highway and drove on the southbound I-215 highway for a total of 40 min (see Fig. 3 for details about the route).

During the first 10 min of the drive, participants were left free to familiarize with the HMI of ACC and LKAS. After the 10-minute familiarization phase, participants were instructed to follow a lead vehicle at a constant distance using ACC. The lead

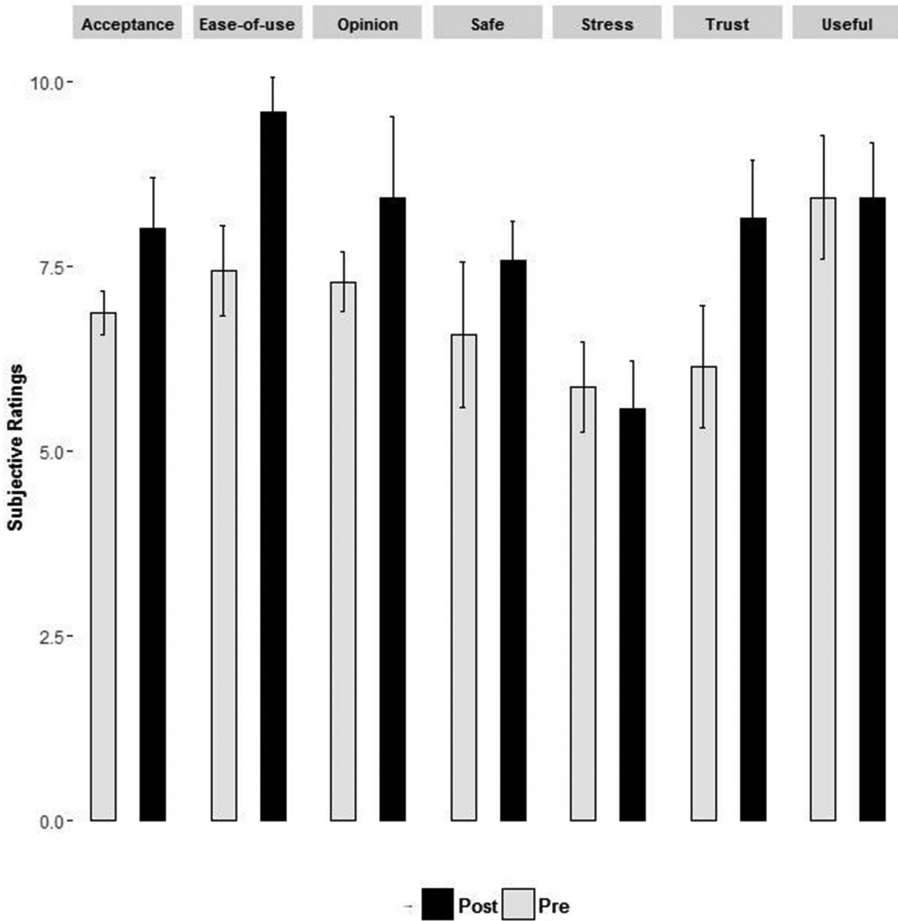


Fig. 4. Subjective ratings for ACC recorded in pre-post questionnaires. Error bars represent standard errors.

For LKAS, significant differences between pre-post questionnaires were found for the stress scale, $t(9) = 2.7, p < .05$, with ratings of stress increasing over time. No significant differences were found for opinion, trust, acceptance, usability, ease-of-use, feeling of safety. Data are presented in Fig. 5.

Participants' opinions and comments toward ACC and LKAS collected via the thinking aloud technique were transcribed. A list of selected comments is presented in Table 1.

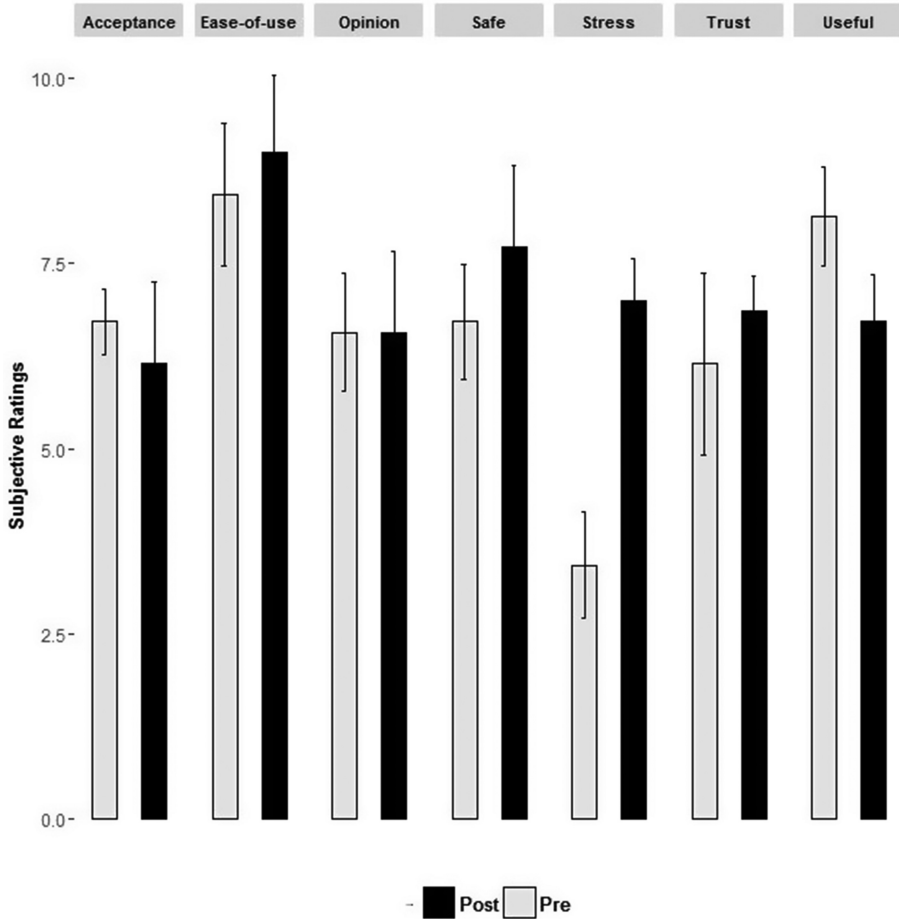


Fig. 5. Subjective ratings for LKAS recorded in pre-post questionnaires. Error bars represent standard errors.

Sentiment analysis [9, 10] was run on the transcribed comments to investigate users’ affective state toward ACC and LKAS. The RSentiment package (version 2.1.1 [11]) was used for this analysis. Sentiment indexes were calculated and presented in Table 2. Positive indexes are indicative of more positive affective states whilst negative indexes are indicative of more negative affective states.

Table 1. Selected list of comments

Positive	Negative
<i>Adaptive cruise control</i>	
“I feel way more comfortable with it Before I was tense and ready to hit the brakes at any minute but now I’m fine I just don’t think that it matches the way other drivers drive though”	“It makes me anxious because the car is slow in accelerating in response to the lead car accelerating”
“The car slowed down nicely. It was not jarring”	“I feel more nervous in higher traffic situations like there is too much going on [to trust the vehicle]”
“As a person that uses cruise control a lot I would definitely use ACC it would be really nice”	
“I still feel like this is a safe distance from the vehicle in front”	
<i>Lane keeping assist system</i>	
“I feel like it’s subtle enough that I still feel like I’m in control of the vehicle”	“I feel like I’m zoning out a little bit more”
“I feel even more comfortable, I almost forget that it’s on”	“I don’t like that I can feel it pushing back”
	“I feel like I would trust this more at a slower speed”
	“I’m a little skeptical, to be honest. It’s not as centered as I would drive myself, and that throws me off. So then I have this internal debate, do I trust the machine more than myself?”
	“I still don’t like it. I don’t trust it at all, no way, especially after those curves”

Table 2. Sentiment indexes calculated for LKAS and ACC. Average and standard deviations are presented.

ADAS	Average	Standard deviation
ACC	0.45	0.34
LKAS	0.23	0.34

4 Conclusion

User trust and acceptance represent aspects of primary importance to be accounted for in the development of human-machine interfaces for autonomous vehicles [7]. One main limitation of the studies investigating *would-be* users’ attitude toward partial to full autonomous driving stems from their indirect experience with such technology. To investigate user interaction with partially autonomous vehicles, in this exploratory study 10 naïve participants drove a SAE level-2 vehicle while engaging Adaptive Cruise Control and Lane Keeping Assist Systems.

Compared to the ratings recorded in the pre-questionnaire, results show that ratings of trust and ease-of-use toward ACC increased over time. This suggests that participants with no prior experience with Adaptive Cruise Control trusted this system more and found it easier-to-use over time. Such findings are in agreement with the study of Kazi et al. [8] in which participants drove a simulated vehicle equipped with ACC with different levels of reliability: 0% (ACC completely unreliable), 50% (ACC reliable half of the times), 100% (ACC always reliable). As the system became more reliable, ratings of trust increased as a consequence and reached ceiling levels after 7 days of exposure to the system.

Different results were found for LKAS. A significant difference in the ratings of stress was found in this study, with ratings of perceived stress increasing in the post-questionnaire compared to those in the pre-questionnaire. This suggests that engaging and interacting with LKAS during highway driving was perceived as a more difficult task compared to interacting with ACC. Such pattern of results is supported by sentiment analysis data suggesting that participants tended to use words with more positive connotations to describe their interaction with ACC compared to those used for LKAS. In particular, participants did not like when LKAS maintained the vehicle in a position within the lane different from that they would maintain during highway driving. Further, some participants suggested that they could possibly trust the system more at slower speeds. Such difference in results between LKAS and ACC may be explained by the fact that, although participants did not have prior experience with ACC, all of them had prior exposure to more archaic speed control systems, i.e., standard cruise control. This might have caused them to be more accepting of ACC and, thus, more skeptical toward LKAS, a system they were mostly unfamiliar with.

This study is of the main importance for addressing future challenges with highly and fully-autonomous vehicles. Results showed that participants, although being trustful of adaptive speed control systems, found interacting with lane maintenance systems a stressful task, especially given the difference between their “driving style” and that operated by the system. Future research is therefore needed to design adaptive, collaborative assistance systems capable of operating the vehicle in ways that are easier for users to trust.

Acknowledgments. This study was funded by the AAA Foundation for Traffic Safety.

References

1. SAE: Automated Driving. Levels of Driving Automation are Defined in New SAE International Standard J3016 (2014). http://www.sae.org/misc/pdfs/automated_driving.pdf
2. ERTRAC: Automated Driving Roadmap (2015). <http://www.Ertrac.Org>, 46
3. Itoh, M., Horikome, T., Inagaki, T.: Effectiveness and driver acceptance of a semi-autonomous forward obstacle collision avoidance system. *Applied ergonomics* **44**(5), 756–763 (2013)
4. Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., Nass, C.: Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *Int. J. Interact. Des. Manuf. (IJIDeM)* **9**(4), 269–275 (2015)

5. McCarty, M., Funkhouser, K., Zadra, J., Drews, F.: Effects of auditory working memory tasks while switching between autonomous and manual driving. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **60**(1), 1741–1745 (2016). SAGE Publications
6. Kyriakidis, et al.: Public opinion on automated driving: results of an international questionnaire among 5000 respondents. *Transp. Res. Part F: Traffic Psychol. Behav.* **32**, 127–140 (2015)
7. Choi, J.K., Ji, Y.G.: Investigating the importance of trust on adopting an autonomous vehicle. *Int. J. Hum. Comput. Interact.* **31**(10), 692–702 (2015)
8. Schoettle, B., Sivak, M.: Public opinion about self-driving vehicles in China, India, Japan, The U.S., The U.K. and Australia, (UMTRI-2014-30), pp. 1–85, October 2014
9. Cambria, E., et al.: New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst.* **28**(2), 15–21 (2013). <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6468032>
10. Bellare, M., Rogaway, P.: The exact security of digital signatures-how to sign with RSA and rabin. In: Maurer, U. (ed.) *EUROCRYPT 1996*. LNCS, vol. 1070, pp. 399–416. Springer, Heidelberg (1996). doi:10.1007/3-540-68339-9_34
11. RSentiment: Package ‘RSentiment’ (2017). <https://cran.r-project.org/web/packages/RSentiment/RSentiment.pdf>

The Human Element in Autonomous Vehicles

Jerone Dunbar^(✉) and Juan E. Gilbert

University of Florida, Florida, USA
{jerone, juan}@ufl.edu

Abstract. Autonomous vehicle research has been prevalent for well over a decade but only recently has there been a small amount of research conducted on the human interaction that occurs in autonomous vehicles. Although functional software and sensor technology is essential for safe operation, which has been the main focus of autonomous vehicle research, handling all elements of human interaction is also a very salient aspect of their success. This paper will provide an overview of the importance of human vehicle interaction in autonomous vehicles, while considering relevant related factors that are likely to impact adoption. Particular attention will be given to prior research conducted on germane areas relating to control in the automobile, in addition to the different elements that are expected to affect the likelihood of success for these vehicles initially developed for human operation. This paper will also include a discussion of the limited research conducted to consider interactions with humans and the current state of published functioning software and sensor technology that exists.

Keywords: Autonomous car · Autonomous vehicle · Connected car · Driverless car · Human vehicle interaction · Self-driving car

1 Introduction

Automotive and technology companies have been exploring the opportunities focused on providing consumers with a fully autonomous vehicle. Delivering such a vehicle for consumers has been identified as one of the major challenges in Computer Science [1, 2]. Related terms have been used to describe autonomous vehicles, such as self-driving car, driverless car, driverless vehicle and autonomous car; however, for the purposes of this paper an “autonomous vehicle” is one capable of performing one or more driving related tasks independently [3]. It is also important to note here that unless “fully” or “100%” autonomous is explicitly mentioned, an autonomous vehicle in the context of this survey is a vehicle with one or more of these automated, semi-autonomous or self-driving features. Human vehicle interaction for autonomous vehicles in the context of this paper is centered on the human interaction with private passenger vehicles that may or may not require human supervision. Human interaction with buses, trucks and motor bikes will be briefly discussed, but the focus is primarily on passenger cars. Automakers and technology companies are working to deliver a fully autonomous vehicle available for purchase to consumers. Companies that have publicly announced their intentions on doing research in this space include Google, Mercedes, BMW, Nissan, Volkswagen, Audi and Volvo [4, 5]. Various automotive and technology companies are racing to be the first to deliver an autonomous vehicle to their customers that can operate on all roads

[6, 7]. Some automotive companies have publicly stated that they will be able to deliver an autonomous vehicle to consumers as early as the year 2020 [8, 9].

This survey will provide an overview of autonomous vehicles, their current state and implications. There will be a major emphasis on the importance of human vehicle interaction and control delegation in autonomous vehicles based on researchers working in this area. While human interaction in the vehicle is a vital component, this survey will also explore many other areas that are related to or likely to affect human beings based on current research in regards to autonomous vehicles. Section 1.1 of this survey will provide a general discussion around the autonomous vehicle space, advantages and disadvantages of autonomous vehicles, growth in recent years, different levels, evolution of advanced driver assistance systems and the importance of human vehicular interaction in autonomous vehicles. Section 2 provides a discussion around the current state of autonomous vehicle technology development, the eight most pressing areas related to autonomous vehicles in the literature and the lack of focus on user experience for autonomous vehicles. Section 3 discusses literature related to human-interaction and control in flight automated systems and potential areas of learning for researchers working in the autonomous vehicle space. Section 4 provides a general discussion of the suggestions going forward considering the areas related to autonomous vehicle development. Section 5 summarizes and concludes the survey.

1.1 Advantages and Disadvantages of Autonomous Vehicles

One of the major reasons behind having autonomous vehicles discussed in the literature is the emphasis on safety and their potential to significantly reduce traffic accidents that typically would have been caused by human error [3, 10–12]. Improving overall roadway safety by reducing traffic accidents has been identified as one of the biggest motivators for the development of autonomous vehicles [3, 13, 14]. According to research conducted by Klauer et al., driver inattention has been identified as the cause of almost 80% of motor vehicle accidents [15]. Driver inattention includes the driver engaging in secondary tasks, driver drowsiness, driving-related attention from the forward roadway or non-specific eye glance away from the forward roadway [15]. Many of these distraction related accidents are expected to be eliminated by the implementation of autonomous vehicles, since they do not get distracted, make significantly less errors and do not get drowsy, compared to human beings [3, 11, 14]. Another key advantage to the development of autonomous vehicles is that they would appear to be better equipped to endure the long trips that are monotonous or tiresome for human drivers [16]. On the contrary, there are many factors that may be considered negative outcomes or disadvantages resulting from the development of autonomous vehicles. One possible negative outcome, especially for driving enthusiasts, is that human-driving may eventually become illegal. Tesla's Chief Executive Officer, Elon Musk is one of the many supporters of making the operation of traditionally human-driven vehicles banned once autonomous and self-driving vehicles become widely used [17]. Various other researchers believe that human driving will eventually become illegal [9, 18]. Another possible negative impact of autonomous vehicles is that the level of expertise associated with adult drivers may decline and people may

eventually become bad drivers due to the lack of actual human-controlled driving experience [19]. Loss of driving skill is likely to be a problem considering autonomous vehicles cannot independently operate on all roads and the vehicle will therefore need a human driver whenever there is a malfunction or system limitation [11, 20, 21]. Research conducted by Lu and Winter suggests that before fully autonomous vehicles are on the roadways, humans will need to supervise their automated cars [5]. Trust in the technology may become too high, security flaws related to stored driver information and personal data may increase in risk, and a vast majority of the people currently working in the public transportation sector would no longer have a job [19, 22]. It is important to note here that the loss of driving skill will not be an issue when there are only fully autonomous vehicles on roadways; however, this is likely to be an area of concern once human control is expected or required while driving [11, 20]. In other words, when the vehicle needs to return control to drivers who become too relaxed in a vehicle with automated features, driving skills will diminish as suggested by prior research [19]. Loss of driving skill is likely to be problematic on our roadways in the future when the vehicle needs to return control to human drivers who no longer remember important driving skills for safe vehicle operation [11, 20]. Subsequent sections will discuss in greater detail how other factors may affect the driving experience such as decreased situational awareness and increased cognitive workload among many others. There are many other negative consequences but these aforementioned factors are a few that have a direct impact on humans based on current research. Many factors also exist that can affect human beings, which may hinder adoption, such as legal implications, initial high cost of the technology, network or infrastructure security, changes in infrastructure, and time required for widespread adoption of autonomous vehicles [3, 9, 11, 12]. These factors will be discussed in greater detail in subsequent sections.

1.2 Source of Rapid Growth in Autonomous Vehicle Research and Development

The Defense Advanced Research Projects Agency (DARPA) is the acclaimed United States federal agency that explores seemingly impossible capabilities for new technologies [3]. DARPA hosted the first Grand Challenge in 2004 [3]. DARPA Grand Challenge participants aimed to develop an autonomous vehicle that was able to successfully navigate desert trails and roads at high speeds, which brought significant attention to autonomous vehicle development. Numerous vehicles were created by a variety of companies and universities to participate in the challenge. No vehicle was able to complete the challenge that year; however, in the following year (2005) after extensive research, a handful of vehicles were able to successfully complete the challenge [3]. Since some vehicles were actually able to complete the challenge in 2005, this demonstrated the potential for additional research in this area and hopes for autonomous vehicle researchers. DARPA later organized the Urban Challenge that took place in 2007. This was the first major large scale challenge where autonomous vehicles would need to prove themselves capable of handling a vast majority of scenarios from an urban setting as well as being able to interact with other moving



Fig. 1. Urban challenge winner 2007 from [3].

vehicles while obeying the rules of the road [3]. The Tartan racing team, which consisted of individuals from Carnegie Mellon University, General Motors, Caterpillar, Continental and Intel, won the Urban Challenge with their autonomous vehicle called “Boss”, pictured in Fig. 1 below [3].

Even though the Urban Challenge was valuable for research and continuous work in this area, it had many limitations. Some of these were noted by Urmson et al., such as no pedestrians, no varied weather, and no dense traffic [3]. Other factors contributing to a potentially ungeneralizable setting included no traffic lights, only low speed testing (under 35 mph), no animals, no bicyclists or skateboarders, and only a limited subset of the rules outlined by the Department of Motor Vehicles among many others [11]. The Urban Challenge was groundbreaking but it really only had a subset of the complexity of situations that could occur in real driving environments. In spite of its limitations, the DARPA Grand challenge led to recent developments by Google on the “self-driving car” [8]. The popular work by Google on self-driving cars has brought greater attention to the feasibility of autonomous vehicles from automobile manufacturers, technology companies and other agencies building future autonomous vehicles.

1.3 Levels of Vehicle Automation

The different levels of vehicular automation have contributed to both the growth and concerns relating to autonomous vehicles [23, 24]. At the time of this publication, there is no fully autonomous vehicle available for purchase to the general public that can independently operate on all roads. However, there are vehicles with Advanced Driver Assistance Systems (ADAS) currently available for purchase that permit a driver to operate the vehicle in specific circumstances without continuous and direct human input. A Tesla is an example of such a vehicle that can temporarily control the powertrain, brake and steering via the Tesla AutoPilot feature, but carries a starting price of around \$70,000 [25]. A Tesla may be too expensive for most Americans, considering top selling vehicles in the United States cost less than \$27,000 [26]. According to the National Highway Traffic Safety Administration (NHTSA), there are five specific levels to the

Table 1. NHTSA autonomous vehicle levels [27].

No-Automation (Level 0)	The driver is in complete control of the primary vehicle controls
Function-specific automation (Level 1)	This involves one or more specific control functions that are independently automated, such as electronic stability control pre-charged brakes, where the vehicle automatically assists with braking to enable the driver to regain control of the vehicle or stop faster than driving alone
Combined function automation (Level 2)	Involves automation of at least two primary controls (steering, powertrain and/or brakes) functions designed to work in unison to relieve the driver of control of those functions. Lane keeping assist paired with adaptive cruise control is an example of this automation level. The driver remains fully responsible for monitoring the roadway. The automated system may need to return control to the driver with very little to no warning
Limited self-driving automation (Level 3)	Includes vehicles that allow the driver to cede full control of all safety-critical functions under certain traffic or environmental conditions to rely heavily on the vehicle to monitor for changes in those conditions eventually requiring transition back to the driver for control. The driver is expected to be available for occasional control, but with sufficiently comfortable transition time. Example is the Google self-driving car
Full self-driving automation (Level 4)	Is a vehicle designed to perform all safety-critical driving functions and monitor roadway conditions for an entire trip. The driver is not expected to be in control at any point during the trip. To date, this vehicle does not exist

different types of vehicle automation [27]. NHTSA is a part of the Executive Branch of the United States government and is also part of the Department of Transportation [27]. These automation levels created by NHTSA will be referenced throughout the entirety of this paper, since it is the governing body for transportation on roadways in the United States. These automation levels are outlined in Table 1 below:

The Society of Automotive Engineers (SAE) is a United States based global organization whose vision is to be a “leader in connecting and educating engineers while promoting, developing and advancing aerospace, commercial vehicle and automotive engineering” [28]. SAE has also defined levels of vehicular automation, which is similar but not exactly the same as NHTSA. Although these automation levels are different, NHTSA publishes some of their work in the SAE technical conferences [29]. Please see Fig. 2 below, which shows a chart of the SAE automation levels from no automation to full automation. These SAE automation levels are worth noting considering their global recognition [28] and may add more clarity to these different levels from an autonomous vehicle standpoint. It is also worth mentioning that SAE cannot directly set or change laws, as NHTSA is able to [30], therefore the SAE automation levels will not be used extensively in this paper.

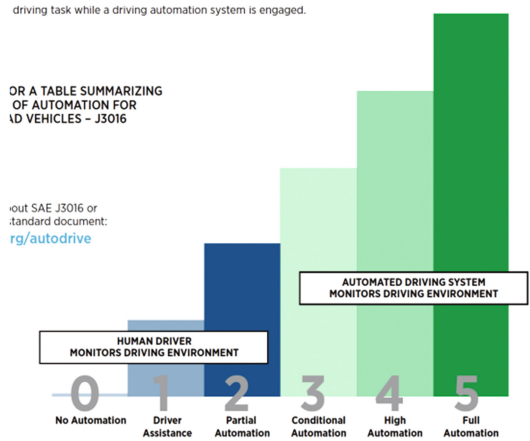


Fig. 2. SAE Levels of automation for on-road vehicles [31].

Please see Table 2 for details on what each numerical representation of the automation levels means, as defined by SAE. Again, these do not map exactly to the levels outline by NHSTA. It is also very apparent that NHTSA has 5 levels (0–4), while SAE has 6 levels (0–5).

Table 2. SAE narrative for automation levels [31].

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/Deceleration	Monitoring of Driving Environment	Fallback Performance of Dynamic Driving Task	System Capability (Driving Modes)
Human driver monitors the driving environment						
0	No Automation	the full-time performance by the human driver of all aspects of the dynamic driving task, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	Driver Assistance	the driving mode-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the human driver perform all remaining aspects of the dynamic driving task	Human driver and system	Human driver	Human driver	Some driving modes
2	Partial Automation	the driving mode-specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the human driver perform all remaining aspects of the dynamic driving task	System	Human driver	Human driver	Some driving modes
Automated driving system ("system") monitors the driving environment						
3	Conditional Automation	the driving mode-specific performance by an automated driving system of all aspects of the dynamic driving task with the expectation that the human driver will respond appropriately to a request to intervene	System	System	Human driver	Some driving modes
4	High Automation	the driving mode-specific performance by an automated driving system of all aspects of the dynamic driving task; even if a human driver does not respond appropriately to a request to intervene	System	System	System	Some driving modes
5	Full Automation	the full-time performance by an automated driving system of all aspects of the dynamic driving task under all roadway and environmental conditions that can be managed by a human driver	System	System	System	All driving modes

1.4 Automobile Evolution and Advanced Driver Assistance Systems

In addition to specific levels of automation, ADAS have evolved significantly over the years from single independent systems such as Anti-Lock Braking (ABS) from the 1970s to more advanced multi feature systems today such as lane keeping, blind spot assist or adaptive cruise control [9–11]. There has been a plethora of driver assistance systems that have been released to contribute to the progression of ADAS, and the number continues to increase [9, 14]. While ADAS generally focuses on specific or a single advanced driving technology, NHTSA has outlined automation levels that center on a combination of these features and eventually full automation. Even though ADAS has undoubtedly been around before any autonomous vehicle, the work within the ADAS space has played a salient role for autonomous vehicle development, therefore an autonomous vehicle is essentially a vehicle with a plethora of ADAS features. In the past decade some the most advanced driving assistance features include single lane highway semi-autonomous driving, blind spot detection, surround view systems, park assist, forward collision warning systems, lane departure warning/lane keep assist and many others [32]. While none of these features allow the vehicle to independently operate without a human driver, some of the individual components can be used in the development of fully autonomous vehicles. For example, electronic stability control (ESC) is a relatively old ADAS feature dating back to 1987 [32]; however, it is likely that an autonomous vehicle will have such a feature or something similar that will continuously monitor steering and vehicle direction and intervene when traction with the roadway is not consistent or when skidding begins to occur [33]. Some of these individual ADAS components can help researchers working on autonomous vehicles in the sense that all parts of a fully autonomous vehicle will not need to be built from scratch and much can be learned from the ADAS technologies that already exists.

1.5 Importance of Human Vehicle Interaction

As mentioned previously, autonomous vehicles will not initially be able to handle all driving scenarios and therefore circumstances exist where control of the vehicle will need to be returned to the driver [20, 34, 35]. The literature suggests that certain operational conditions are problematic for autonomous vehicles without human input, such as construction zones, areas where an accident has recently occurred, approaching vehicles with a rare appearance, unstable road situations such as snow, detecting known objects at speeds over 81 mph, unknown objects, ice or potholes and other unexpected situations [8, 12–14]. This transition of going from the fully autonomous driving experience back to human-controlled manual driving has been noted as one of the major challenges for the producers of the fully autonomous car of tomorrow [5, 15, 16]. Human vehicle interaction in regards to human-controlled manual driving is likely to remain of major importance and a challenge until fully autonomous vehicles are able to drive on all roads [5, 20, 34, 35]. Simply put, they would need to be completely reliable without any human input, defined as a level 4 fully autonomous vehicle by NHTSA [16, 17]. Human-Computer Interaction (HCI) researchers, designers and automotive manufactures have an arduous task ahead to ensure that this human to

machine experience is carefully crafted [5, 22]. Essentially, much will need to be in place to best support autonomous vehicles compared to the driving environment today, such as vehicle-to-infrastructure communication, vehicle-to-vehicle communication, policy changes, and new ways for human interaction with autonomous vehicles among others [9, 22]. The human to vehicle interaction piece is one of the most related areas to this survey paper and it has been noted in the literature that very little is known about designing interfaces for autonomous vehicles [22, 39]. The interfaces relevant to this survey in the context of autonomous vehicles will be discussed in subsequent sections. Autonomous vehicles will need the ability to recognize all possible objects or things that could be on or alongside the roadway before they can operate as truly autonomous [9, 27]. It also important to mention that even though there has been years of work on autonomous vehicles centered on the functionality of these systems, little attention has been given to the human interaction that will occur with these vehicles that humans are not familiar with [3, 5, 22]. If enough attention is not paid to human interaction, research has suggested that this could lead to unfavorable circumstances, such as mode confusion, user distrust or overreliance in automated systems, which are likely to lead to accidents [22, 40, 41].

Another salient area related to human vehicle interaction centers on how ethical situations will be handled by autonomous vehicles and how blame is assigned in the event of an accident [42–44]. Some open research questions noted in the literature by these researchers include: “what is the correct way to program an autonomous car for ethical situations? Does it matter if the potential individual(s) in a crash are adults or children? Should age be a deciding factor for survival?” [42–44]. Questions regarding liability in the event of an accident will also arise such as: “is it the fault of the car owner, vehicle manufacturer or even programmer who worked on the software” [45]. These are related to human interaction in the sense that it is unknown how much control the driver will have in these situations or if the driver will be allowed to interact with the vehicle or have a say in a potential traffic accident, based on current literature [42–44]. Although ethical situations are related to human interaction in the vehicle, the topic of ethics will be discussed in later sections.

Take-over request (TOR) is an imperative factor as it relates to human interaction with autonomous vehicles. Researchers are currently working to identify precisely when a TOR needs to be issued prior to some limitation in automation or some unexpected circumstance that the vehicle cannot handle [20, 34, 35]. Much work is still needed in this area as there is a wide array of factors to consider such as other tasks in which the driver is engaged (which is potentially a long list, as automation improvements permit drivers to look away from the forward roadway for longer periods), the speed of the vehicle, and the amount of roadway available for correction, among many others [20, 34]. A pictorial representation of a TOR is demonstrated in Fig. 3 below to provide a better understanding of the complexity of the scenario.

Prior research has demonstrated that drivers currently indulge in a plethora of distracting activities while driving such as engaging in cell phone use and interacting with the built-in information system [22]. Within these two main tasks alone, drivers perform numerous sub activities such as texting, talking, using social media, and/or adjusting the climate or radio, among many others that distract them from the immediate driving task at hand [46, 47]. Accident data suggests that people currently do

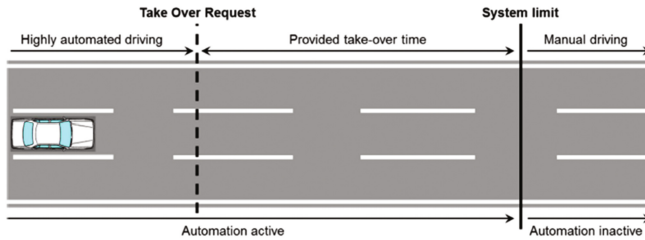


Fig. 3. Illustrative example of a TOR from [20].

more than they can handle while driving on roads today, which is evident from the 3,179 people killed and 431,000 that were injured in 2014 in the United States alone [48]. According to the NHTSA, these deaths and injuries were a direct result of distracted driving [48]. It is clear that there are distraction concerns in cars today and this is likely to increase as more and more driver assistance and autonomous features are included in automobiles. The possible combination of tasks in which the driver can engage could become even longer considering the gamut of activities drivers perform or engage with while driving. Some general examples include using a navigation system (in-vehicle or mobile), searching for an item in the car, eating a snack/meal, drinking a beverage, etc. Automation has assisted drivers in becoming safer and helping to reduce accidents, but they also make it easier for the driver to engage in secondary activities [24]. When taking into account the increase in vehicular automotive and driver assistance features in recent years, engaging in other activities while driving may become less difficult.

Take-over request (TOR) is of utmost importance here and handing over control to the driver has been identified as one of the most daunting tasks for HCI researchers, designers, and automotive manufacturers [14, 15, 17, 18, 22]. Drivers traveling at different speeds may require a different mode or process to transition back to manual driving, especially considering that the time required for the driver to take control may vary depending on the driver and/or situation. One major research challenge lies in identifying how much time a driver needs to regain control of the vehicle safely [3, 14]. Two additional factors potentially germane to this transition include the personality of the driver as well as the secondary activities in which the driver is engaged when the vehicle must return control to the driver. Research has suggested that a driver could be viewed as being in different levels of attention, such as monitoring the road ahead, drowsy, sleeping, reading a text message or email, talking on the phone, and talking to a passenger, among many others [46, 50]. Returning control to the driver is only one area that appears to be difficult to account for in all related possible situations. Autonomous driving is affected by many situations that are predictable; however, there are also unforeseeable driving situations [38]. These unforeseeable situations are intimidating due to the fact that, as an unforeseeable circumstance, the unfortunate event would have to occur in order to see the need for and apply the resolution [38]. In order to address many of these unpredictable scenarios, research suggests that continuous testing is needed for robust implementation and a more functional autonomous vehicle [1, 38]. It may be in the best interest of automakers and technology companies

to account for as many of these challenging scenarios as possible through testing, prior to the release of these vehicles [38].

2 Current State and Implications of Autonomous Vehicles

2.1 Autonomous Vehicle Technology

The technology available in cars is tremendously advanced, and when considering the luxury line automakers, their technologies continue to improve. From a computing perspective, vehicles that have autonomous features are primarily dependent on GPS, cameras, laser range finders, radar and extremely accurate maps of the environment [1]. One key technology used in autonomous vehicles is a light detection and ranging (LIDAR) sensor, capable of scanning one million 3D points per second. The widely known Velodyne LIDAR sensor needed for fully autonomous vehicle operation costs between \$30,000 and \$85,000 for the sensor alone, which is still considerably expensive for the average consumer [3, 8, 23]. Figure 4 below provides a visual representation of the unprecedented Google self-driving car, with the Velodyne LIDAR mounted on the roof of the vehicle.



Fig. 4. Google self-driving car from [9].

To date there is no known precise combination of sensors, cameras, LIDARs and other technologies that are required to be included in all autonomous vehicles. Outside of information disclosed in a patent that a company specifically owns, novel information regarding product development details is not usually disclosed to the public as with many other new technologies [52, 53]. Google has already demonstrated capabilities of an autonomous vehicle that can drive, but as noted earlier a level 4 NHTSA vehicle that does not require any human input does not yet exist [27]. In terms of what is currently known in literature about the technologies in autonomous vehicles, they have dedicated systems for motion planning (trajectory generation, on-road navigation and zone navigation), a perception system responsible for providing a model of the world to the behavioral and motion planning subsystems (moving obstacle detection

and tracking, static obstacle detection and mapping, roadmap localization and road shape estimation), mission planning (detection blockages, handling blockages), behavioral reasoning (intersections and yielding, distance keeping and merge planning, and error recovery), software infrastructure (communications library, interfaces library, configuration library, task library, debug logger and log/playback) and testing that is sometimes intertwined with the software stack [3]. Pink et al. provided a great illustration of the current main sensor technology in autonomous vehicles, in Fig. 5 below [38]. This illustration is not intended to be exhaustive; however, it provides an overview of the sensing technologies included in these vehicles. This image depicts the sensing technologies employed specifically by Bosch autonomous vehicle research division and is unlikely to be exactly the same for other companies working on autonomous vehicles. The technologies used in autonomous vehicles will vary to some degree dependent on the automotive manufacturer or technology company and the level of autonomy that the vehicle can support.

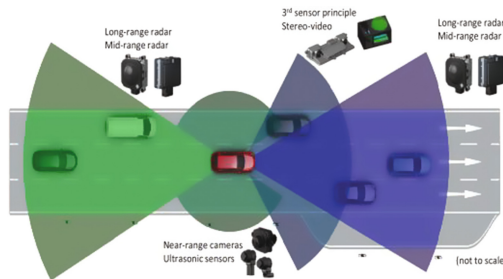


Fig. 5. Field of view of sensors for autonomous vehicles from [38].

2.2 Implications of Autonomous Vehicles

Legal Implications. Concerns around legality and responsibility are of monumental importance since accidents are likely to occur in the future where an autonomous vehicle is not being directly controlled by the driver [44]. The question is not whether accidents with autonomous vehicles will occur, but a matter of when. There have been two recent deaths when Tesla's Autopilot was engaged in addition to other accidents not resulting in deaths where AutoPilot malfunctioned. Tesla's Autopilot feature has the ability to temporarily control the vehicle's powertrain, brake and steering under specific highway conditions [25]. On May 7th 2016, there was an accident that resulted in the death of a driver using Tesla's AutoPilot [54]. AutoPilot is especially relevant here since there may be direct implications on future autonomous vehicles stemming from this particular accident. In this and similar future scenarios, some open questions include: Who should be held responsible? Would it be Tesla, the dead driver of the car, the truck that collided with the Tesla, the owner of the car, or someone else? These malfunctions are increasing as advanced driver assistance features (with semi-autonomous capabilities) become prevalent. More recently, on September 14th 2016, the accident details were released of a Tesla with AutoPilot engaged that resulted in the

death of a 23-year-old man [55]. The accident details by The Drive news agency included video footage showing the Tesla colliding head on with a street-cleaning truck that was stopped on the side of the road [55]. The video footage from the accident suggests that the Tesla involved in the accident drove into the street-cleaning truck at constant speed and did not appear to slow down [55]. AutoPilot malfunctions are becoming a growing problem for Tesla and potentially the entire automotive and/or technology industry. Two similar accidents (fortunately with no human deaths) have recently occurred with Tesla's AutoPilot engaged where the Tesla in question failed to notice a vehicle that was stopped on the side of the road in front of it [55–57]. The release of the news about recent Tesla accidents does inform the public about a potential system failure of Tesla's AutoPilot automated feature. The exact impact of the release of this information on public perception of autonomous vehicles is unknown. How frequently these accidents occur is likely to be a factor, since when accidents are infrequent, people may not consider these accidents with automated cars as a major problem. On the contrary if many accidents frequently occur with automated vehicles malfunctioning, this may lead to a negative association by the public with automated vehicles. Additionally, if accidents do occur but are not released to the public then automated crashes may appear as less of a problem to the public. Accidents similar to these where an automated system malfunctions could lead to some customers being concerned about the functionality of autonomous technologies, while it may lead others to completely lose faith or interest in these technologies [58]. It is unlikely that customers will be inclined to purchase a vehicle that is known to have automated features that malfunction if these accidents continue to occur. These accidents should certainly not be taken lightly by technology companies and automakers as vehicles with autonomous features are still in sensitive, growing and developing stages.

The discussion to address questions regarding liability becomes even more sensitive and convoluted depending on how all the scenarios regarding control are formalized [43, 44]. For example, when a human driver identifies danger or a forthcoming accident, action is typically taken to prevent or avoid the problem. If the vehicle is not designed to return control to the driver or the outcome of the vehicle's pre-programmed solution to the identified problem is one that the driver disagrees with—who should be held responsible? [43, 44]. The outcome refers to the multiple possibilities that an autonomous vehicle can select from for accident-prone situations. Regulators will need to account for this and all other possible scenarios of liability with autonomous vehicles.

Infrastructure Implications. The infrastructure for autonomous vehicles is related to humans in the sense that a robust and efficient infrastructure will lead to a safer driving environment and require little to no input from a human driver in an autonomous vehicle [10, 12]. There has been ongoing work to develop the appropriate vehicle-to-vehicle, and vehicle-to-infrastructure communication capabilities for autonomous vehicles [12]. One essential piece to future autonomous vehicles is Dedicated Short Range Communications (DSRC), which supports communication between vehicles (vehicle-to-vehicle) and also from a vehicle to a communication network of roadside units [12]. DSRC is expected to improve the reliability, safety and

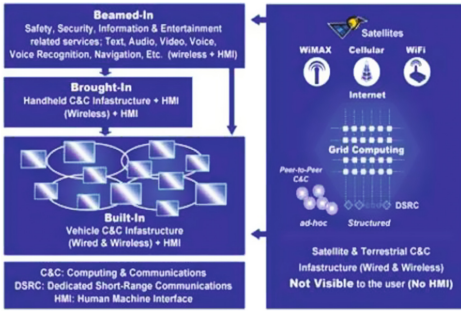


Fig. 6. From [12], illustrates that autonomous vehicles are likely to incorporate beamed in, brought in, DSRC as well as Satellite capabilities in order to support their communication needs.



Fig. 7. Vehicle infrastructure initiative from [12], which is grounded on IEEE 1609.x & IEEE 802.11p standards.

performance of autonomous vehicles [8]. See Fig. 6 from Gharavi et al. below for a mapping of the communication possibilities of automated vehicles.

Figure 7 provides a broader picture of the components and communication channels of the vehicle infrastructure or vehicle-to-infrastructure initiative.

If vehicles are able to communicate with each other via vehicle-to-vehicle and vehicle-to-infrastructure communication capabilities, then they would be able to better inform each other of road conditions and hazardous situations [12]. Vehicle-to-infrastructure communication would support the ecosystem of autonomous vehicles and they will be able to seamlessly communicate with emergency services and help keep the maps up to date with added precision [12]. As noted earlier, vehicle-to-vehicle communication and vehicle-to-infrastructure communication will reduce driver stress and lead to a safer driving environment [10, 12]. Reduced driver stress can eventually lead to a driving environment where humans could focus on other activities in the car and enjoy other things on their trip outside of driving. Researchers working in this space have noted that necessary physical infrastructure changes will be needed to permit autonomous vehicles to communicate seamlessly with the infrastructure and with each other [9, 10, 12].

General Implications on Trucks, Buses and Motor Bikes. Autonomous vehicles are also likely to impact human life for people working directly in the transportation sector. As it was briefly mentioned previously, many working people in the transportation sector will be directly affected by fully autonomous vehicles. In the United States alone, there are over 3 million truck drivers [59]. Fully autonomous vehicles will lead to the elimination of jobs for truck drivers, if a human driver is not required to be present [60]. According to the United States Department of Labor there are almost one million bus and taxi drivers [30, 31], whose jobs will be eliminated once fully autonomous vehicles are available. These implications are clearly not directly focused on human interaction, but the general implication of autonomous vehicles will potentially change or affect the lives of millions of people.

It is important to note here that while this article is focused on human-interaction with autonomous vehicles, the interaction that people have with varying levels of automation in a personal automobile may be very different for motorbikes, buses and trucks. In reference to buses and trucks, this may potentially be an extremely sensitive area from a design and development perspective considering an accident is likely to be more catastrophic due to their large size and weight compared to a personal passenger car. Additionally, a motorbike operates differently from a truck, bus and even passenger cars. These subtle but important differences will need careful consideration in reference to their design and implementation as these autonomous cars, motorbikes, buses and trucks are created.

Trust. Aeberhard et al. suggest that trust is extremely important in reference to situations when human beings interact with automated systems; these systems should work as expected and consistently work well [21]. If the needs and expectations of the driver are not met, this could have extremely negative effects on trust in automated systems and eventually autonomous vehicles [63]. In terms of trust, both too much of it and too little of it can be potentially harmful [22, 58]. Too little trust in the system will leave drivers on edge all the time about the decisions that the car makes and too much trust in the system may cultivate drivers that may delay responding or orienting themselves back to the driving environment when necessary [15, 26–28]. Finding this middle ground between highly but not fully automated is exceedingly challenging as automakers want drivers to utilize highly automated features [24]. Still, they do not want complete disorientation from the driving environment since vehicles are not yet 100% autonomous. Automakers and technology companies developing the autonomous vehicle of tomorrow are likely to be most concerned about the initial trust in the system since this is related to how much profit they will be able to make, their success and overall acceptance. In other words, if people do not trust automation in vehicles then it is unlikely that they will be willing to purchase an autonomous vehicle. Low trust in automation could then lead to a lower adoption rate for autonomous vehicles. Automakers and technology companies would then need to focus their efforts on methods to increase driver trust. Researchers who have done work with trust in automation have identified training to be an essential component of improving user trust [63]. However, too much trust in automation can again lead to overuse or misuse of automated systems [22, 58, 63].

Privacy. As data is stored in an increasing number of locations and on a wide range of devices, privacy is inevitably becoming a growing concern [67, 68]. Drivers may be concerned about the data their autonomous vehicle collects about them and also who has access to their information [9, 67, 68]. A nearly endless list of nefarious activities could occur with the data captured by connected or autonomous vehicles. These activities include but are not limited to providing incorrect information to drivers, limiting the functionality the driver has, having access to all details regarding past, present and future driver routes, acting as a different vehicle or making use of denial-of-service attacks to take down the network [9, 67]. Knowing the driving habits of a driver could be exploited for marketing, law enforcement or surveillance [68]. Particular attention will be needed on the topic of privacy as autonomous vehicles become more and more prevalent. Changes to existing structures or legal requirements

may be necessary to allow for the evolving needs of users and/or determine the level of vehicle information that should be disclosed [67].

Security. Security is another salient topic in the autonomous space. The work by Petit and Shladover on cyberattacks is the first known research to explore the vulnerabilities that exist that are specific to automated vehicles [8]. They focused on the potential areas of infiltration for attacks, which were extensive. The vulnerable areas include electronic road signs, machine vision, Global Position System (GPS), in-vehicle devices, acoustic sensor, radar, LIDAR, road, in-vehicle sensors, odometric sensors, electronic devices and maps [8]. Considering that each of these attack surfaces often include subcategories, potential areas of attack are even higher. Unfortunately for the autonomous industry, hackers exposed vulnerabilities in a 2015 Jeep Cherokee by demonstrating their ability to take control of the steering, gas, and brake pedals from a remote location [69]. This vehicle, a level 2 on the NHTSA scale, had only a few automated features. A fully autonomous vehicle includes an even wider array of connectivity features, which potentially opens the door to many more opportunities for hackers. The work done by Petit and Shladover has identified many key areas that affords hackers the opportunity to breach the network of an autonomous vehicle; however, research suggests that much more work is still needed in this area [4, 8].

Pricing. Litman makes it apparent in his 2014 research on predictions relating to autonomous vehicle implementations, that the initial cost of an autonomous vehicle is one of the key challenges to deployment of these vehicles [9]. If the average human being cannot afford an autonomous vehicle, then only affluent people will be able to enjoy the use of these vehicles until prices are reduced. This would be contrary to the initial overall goal of autonomous cars being created for a safer driving environment since only the select affluent few would be safer and not the general public [3, 11]. Currently, \$30,000 could buy a top selling car in the United States [26]. The LIDAR system, needed for detection alone, costs between \$30,000 and \$85,000 [51]. This excludes the cost of the vehicle and the many other components needed for detection in an autonomous vehicle [51]. Based on these calculations a consumer would need approximately \$60,000 or more in order to purchase an autonomous vehicle. A \$60,000 vehicle is likely to be too expensive for most Americans, especially considering the top selling automobiles in the United States range between \$16,000 and \$27,000 [26]. Vehicles such as a Tesla with advanced autonomous NHTSA Level 2 capabilities where the vehicle can temporarily control the powertrain, brake and steering via the AutoPilot feature is likely to be too expensive for average customers, considering its starting price of around \$70,000 [25]. As with many other technologies, prices tend to reduce over time with increased production, however the initial cost may be too high for the average car buyer and there is no guarantee on how soon prices will be reduced. Shchetko also notes that it is unclear when an autonomous vehicle will be affordable enough for the mass car market [51].

Time to Adoption. As noted previously in this paper, automakers and technology companies are aiming to deliver a fully autonomous vehicle by the year 2020 [8]. It is important to note that, according to the technology companies and automobile manufacturers working on this technology, the year 2020 is approximately the earliest time

that such a vehicle could be delivered to consumers. Considering the plethora of related concerns for autonomous vehicles, many of which are discussed in this paper, it is likely that their deployment could take even longer than predicted. The belief that these vehicles will not hit the marketplace as soon as expected is shared among researchers [70]. It is important not to only consider the time to deployment but also the time when autonomous vehicles will be in mass production, which is one of the major factors reducing the purchasing cost for consumers [51]. If only high-income individuals can afford autonomous vehicles, then the impact will evidently be less meaningful to the average consumer.

2.3 Lack of Focus on User Experience for Autonomous Vehicles

It appears that the engineering of autonomous vehicles is on track, however, the understanding of the interaction between the vehicle's actions and driver reactions seems much more ambiguous [19]. Consequently, there will need to be a significant emphasis on the human element in autonomous vehicles, considering all possible interactions for all levels of automation. Since the DARPA Grand and Urban challenges, there has been considerable amounts of work put into the development of the algorithms, functionality and technologies needed for autonomous vehicles; nevertheless, there has been a lack of focus on the user experience and interaction between the driver and the car [1, 3, 9, 14, 34]. Further, the communication infrastructure and current autonomous vehicle technology needed to allow for performance without driver input is not advanced enough for immediate vehicle deployment [10, 12]. Adequate time and attention will need to be paid to the transition stages to higher levels of autonomy and the back and forth interactions between the driver and a fully autonomous vehicle. This transition of control back and forth between driver and vehicle will have to occur for some time due to unexpected situations that the vehicle will be unable to handle or due to some type of system limitation or failure [13, 18]. A greater focus is needed on the driver interaction experience otherwise autonomous vehicles are likely to have a much longer time to adoption.

3 Human-Interaction and Control in Flight Automated Systems

There has been a significant amount of effort in the design of the modern aircraft from a holistic perspective, especially with regard to the human-machine interaction in the airplane [35, 36]. A failure or issue in the cockpit is likely to result in a catastrophe affecting a large amount of people, therefore designers and developers have made great efforts to minimize possible errors or issues. The human-machine interaction in an aircraft is inherently different from that of an automobile; however, equal importance must be given to this interaction within the context of the automobile similarly to that which is given within the context of airplane development [19]. While it is true that the cockpit of an airplane is more complex than an automobile in terms of functionality, the roadway has a more extensive range of unexpected and complicated scenarios as well

as items that could cause a collision [34, 37]. Airplanes follow rather strict Air Traffic Control (ATC) rules and are generally on the lookout for other airplanes [19, 37]. Special instructions also exist when flying low to avoid helicopters and high rising objects [19, 37]. On the roadway, a myriad of potential dangers exist that a driver has to be able to react to at any time such as unexpected behaviors from other drivers, motorbikes, cyclists, pedestrians, animals, potholes, and objects or debris obstructing the forward roadway among many others. These are all salient areas of concern that will need to be accounted for in the design of vehicles with autonomous capabilities. Similar to automation in the vehicle, as previously outlined from NHTSA level 0–4, there are different levels of automation that a pilot can use in an airplane [72]. The pilot can select from and combine different levels of automation.

In aviation, the final control of the automation is dependent on the size of the plane. Automation can override the intentions of the pilot for smaller planes, defined as “hard” automation [19, 74–76]. However, with larger aircrafts, “soft” automation is used, where the intentions of the pilot are not overridden by the automated system [19, 74–76]. The idea behind hard automation is to use technology to prevent or limit human error and therefore will not allow a human operator to override preset limits of the system, even if there is an emergency. Airbus planes (small aircrafts), such as the A320, A330, A340, A380, etc., employ this hard protection system [19]. With this hard protection automation, functions go through two phases if it is originated from a human operator [19]. After the human operator performs an action, the system verifies whether the instructions are within system limits prior to the actual execution of those actions on the aircraft’s control surfaces. While with soft automation design, pilots are granted complete authority to override the automated system [74, 76]. Boeing (large) aircrafts use the soft protection system [19, 74, 76]. Intentions from a human operator in the soft protection system are immediately relayed to the aircraft’s control surfaces. If automation identifies an issue or concern in this soft protection system, it may issue some type of cautionary alert but it will not stop or nullify the intentions of the pilot. In safety critical scenarios soft automation may provide feedback, which may require the pilot to apply more force than usual, but again the automation will not completely stop the intentions of the pilot [19, 74].

These two approaches of soft and hard automation have both advantages and disadvantages in aviation. Aircraft manufactures have adopted completely opposite approaches in practice. From this perspective, it is therefore not clear which is definitely best for vehicle automation. Even though it has been noted that hard automation may cause more human factors issues, there are concerns that also exist with soft automation. The work by Young et al. suggests that we can learn much from aviation in regards to automation [19, 77]. However, it is not a direct mapping in reference to soft or hard automation being the optimal implementation for autonomous vehicles. Both hard and soft automation have been used in the automotive space. Anti-Lock Braking is an example of hard automation in automobiles and Automatic Cruise Control for soft automation. Hard automation in the driving environment will not allow the driver to interact with or control the automation mechanism, while soft automation will provide the driver the opportunity to have ultimate control. The functional implementation of hard and soft automation in the driving environment is very similar to aviation. Many researchers have noted that the driving environment is more complex and also that there

is much more variability in the driving environment than in aviation [34, 39]. Similar to aviation, hard and soft automation, has both advantages and disadvantages in the driving environment. In reference to hard automation in the driving environment, Intelligent Speed Adaptation (ISA), a feature that uses GPS position monitoring and maps of database speed limits has been claimed to be able to reduce all injury accidents by up to 37% [34, 40]. The advantages of such a tool is rather clear as the tool would eliminate speeding. However, a major disadvantage of such a feature is that vehicle imposed speed restrictions [19] could potentially cause accidents (especially where there is only a single lane road for each direction of traffic). It would not make sense to implement ISA in an emergency vehicle, nevertheless for situations where a human is rushing to the hospital in a non-emergency vehicle, ISA could be very problematic. Automatic Cruise Control is a prominent example of soft automation and a key advantage is that manual input from the driver will disengage such a system [19]. The disadvantage of this technology is that prior research has found that many drivers failed to reclaim control of the Automatic Cruise Control system in some emergency situations [80]. Another disadvantage includes reduced awareness of the driving environment, since drivers are much less in sync with the driving tasks when Automatic Cruise Control is active [24]. The fact that both hard and soft automation has both advantages and disadvantages does not help the design and implementation research process for future autonomous systems. Even though soft automation seems promising, an entirely new approach to automation for autonomous vehicles may be warranted. Consequently, all possibilities of interaction with an autonomous vehicle need to be critically examined, as it needs to be safe, while at the same time easy for drivers to use and understand.

Another imperative point to consider is that according to the Federal Aviation Administration (FAA), airline pilots have to go through a minimum of 1500 h of flight training before being eligible to earn a license to fly a commercial aircraft [81]. Obtaining a license to drive in the United States has specific age restrictions based on the state in which the applicant resides, however, most states only require the applicant to pass a vision and written exam as well as a physical driving test [82]. Additionally, the training required for drivers can vary. For example, driver A may practice for one month while driver B may practice for a year prior to taking the driver's test, while all pilots have a minimum of 1500 h of training required to be able to fly and comprehensively understand flight controls [81]. Consequently, it is important that automated systems are easy to use and seamlessly integrated into what the driver expects in the variety of situations that could occur. Confusion in an automobile is likely lead to accidents, fatalities and thus become a barrier to adoption [22, 40, 41].

A potential way to address the human-vehicle interaction issues that may occur with future cars is to create a standardized reporting system similar to what is used in aviation [83]. This would allow drivers to report some of the interaction issues that were not necessarily foreseeable prior to the deployment of those particular autonomous vehicles. This would also help to reduce possible accidents, fatalities and consumer frustration for drivers. This is more centralized as opposed to a breaking news story or report, that notifies the public, technology companies and automotive manufacturers of automated human-interaction issues.

4 Suggestions Going Forward

There has been tremendous growth and progress in the automotive space, especially over the past decade and a half. The DARPA Grand and Urban challenges brought much needed attention to autonomous vehicle research and development, and highlighted many of the potential opportunities [3]. ADAS have also contributed to the continuous advancement in automation as well. Automobiles are not only being built to have lower level autonomous features such as level 2 & 3 NHTSA vehicles, but there is groundbreaking work being done to develop the first level 4 fully autonomous vehicle [4, 5].

While the engineering and functionality of autonomous vehicles have been at the center of attention for research and development, there is still much work needed to address the many challenges in automation for drivers. Future research will need to be focused on all possible areas of human-interaction with these autonomous vehicles. There are specific and salient areas of concern, many of which are delineated in this article, such as approaches to control for automation, TOR, individual differences and distraction, among many other factors. If humans are not able to interact or understand these vehicles appropriately, there may be unfavorable shifts in the overall acceptance and use of autonomous vehicles.

Many challenges are still present for automotive research, especially for the most advanced states of autonomous vehicles. The sensor technology, cameras, algorithms, machine learning and infrastructure are still not at the level of functionality needed for safe and immediate autonomous vehicle deployment [1, 10]. For example, the fully autonomous sensing technologies need to be improved for sensors, LIDAR, cameras, and others [1]. There has not been enough autonomous vehicle driving data for vehicles to predict and assess all possible driving scenarios [10]. The infrastructure is also not currently in place to support vehicle-to-infrastructure and vehicle-to-vehicle communication on all roads [9, 10, 12]. There is much work ahead, not only from a purely Computer Science and Engineering perspective, but also from a Human-Computer Interaction, Human Factors and Design perspective. To date, even though work is being done to achieve this goal, there are no NHSTA level 4 vehicles that exist [27].

5 Summary and Conclusion

Automotive and technology companies clearly have an onerous task ahead, not only to ensure that autonomous vehicles operate appropriately, but also to be able to interact with humans for all driving cases that could potentially occur. Human Factors, HCI and Design researchers have a grand opportunity to explore, research, appropriately design, and test all possible human vehicle interaction scenarios to contribute to the success and potentially increase the likelihood of acceptance for autonomous vehicles.

There are also cultural and regional differences in driving behavior that will need to be accounted for, considering that people drive differently in various parts of the world and even drive on opposite sides of the road in some parts of the world. This paper investigated many of the human vehicle interaction scenarios that will need to be considered in order for autonomous vehicles to be accepted in the marketplace. Consideration was also given

to the changes in the legality around driving behavior that will eventually be needed, as well as the necessary infrastructure needed to support these vehicles.

Autonomous vehicles cannot simply replace human drivers [84]. Automation is shifting driving from actively controlling to a state of monitoring; however, research has suggested that human beings are not good at monitoring [34]. Further, humans tend to increase participation in secondary task with an increase in automation and driver assistance systems as discussed in this paper. The area of secondary task involvement by drivers in vehicles with autonomous capabilities will need extensive research moving forward. Research has also suggested that drivers respond more quickly to visual-auditory information requests than they do to requests that are only visual [14, 22]. Alerting systems in vehicles may need to evolve as the levels of vehicle automation has grown and evolved.

Although human interaction with an airplane and other automated systems is not the same as interacting with an autonomous vehicle as outlined in this survey, a plethora of knowledge can be gained from these interactions with other autonomous systems to be used as a guide or reference point. Driver distraction will continue to be a challenge until autonomous vehicles are able to operate without any human input. The amount of secondary devices that can distract the driver has been increasing, considering the myriad of devices that drivers can bring into the vehicle, not to mention the information rich in-vehicle infotainment systems in cars today. A vast amount of additional research will need to be conducted in order to clearly understand the intentions of drivers and how control will be handled in autonomous vehicles, while at the same time taking into account the variations in driver and personality type. Getting the interaction right the first time is even more important considering that people are not familiar with this technology being introduced. There is an exciting journey ahead for the automakers, technology companies, researchers and legislators to create seamless and safe experiences for drivers in order to promote the growth and broader acceptance of autonomous vehicles.

References

1. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3354–3361 (2012)
2. Peterson, K., Ziglar, J., Rybski, P.E.: Fast feature detection and stochastic parameter estimation of road shape using multiple LIDAR. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, pp. 612–619 (2008)
3. Urmson, C., et al.: Autonomous driving in urban environments: boss and the urban challenge. In: Buehler, M., Iagnemma, K., Singh, S. (eds.) The DARPA Urban Challenge. STAR, vol. 56, pp. 1–59. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-03991-1_1](https://doi.org/10.1007/978-3-642-03991-1_1)
4. Schoettle, B., Sivak, M.: A survey of public opinion about connected vehicles in the U.S., the U.K., and Australia. In: Proceedings of the 2014 International Conference on Connected Vehicles and Expo, ICCVE 2014, pp. 687–692 (2015)
5. Lu, Z., De Winter, J.C.F.: A review and framework of control authority transitions in automated driving. *Procedia Manuf.* **3**, 2510–2517 (2015). AHFE

6. Griffith, E.: Who Will Build the Next Great Car Company? *Fortune* (2016)
7. 33 Corporations Working On Autonomous Vehicles, CB Insights (2016). <https://www.cbinsights.com/blog/autonomous-driverless-vehicles-corporations-list/>
8. Petit, J., Shladover, S.E.: Potential cyberattacks on automated vehicles. *IEEE Trans. Intell. Transp. Syst.* **16**(2), 546–556 (2015)
9. Litman, T.: Autonomous vehicle implementation predictions: implications for transport planning. *Transp. Res. Board Annu. Meet.* **42**, 36–42 (2014)
10. Wei, J., Snider, J.M., Kim, J., Dolan, J.M., Rajkumar, R., Litkouhi, B.: Towards a viable autonomous driving research platform. In: *Intelligent Vehicles Symposium (IV)*. IEEE (2013)
11. Levinson, J., et al.: Towards fully autonomous driving: systems and algorithms. In: *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 163–168 (2011)
12. Gharavi, H., Prasad, K.V., Ioannou, P.: Scanning advanced automobile technology. *Proc. IEEE* **95**(2), 328–333 (2007)
13. Sikkenk, M., Terken, J.: Rules of conduct for autonomous vehicles. In: *International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, vol. 1(1), pp. 19–22 (2015)
14. Sun, Z., Bebis, G., Miller, R.: On-road vehicle detection: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(5), 694–711 (2006)
15. Klauer, S.G., et al.: The impact of driver inattention on near crash/crash risk: an analysis using the 100-car naturalistic driving study data. *Analysis*, 226 (2006)
16. Nordhoff, S.: *Mobility 4.0: Are Consumers Ready To Adopt Google's Self-Driving Car?* (2014)
17. Elon Musk: self-driving cars could lead to ban on human drivers, *The Guardian*
18. Samit, J.: Driving a car will be illegal by 2030, *Wired*
19. Young, M., Stanton, N., Harris, D.: Driving automation: learning from aviation about design philosophies. *Int. J. Veh. Des.* **45**(3), 323–338 (2007)
20. Gold, C., Damböck, D., Bengler, K., Lorenz, L.: Partially automated driving as a fallback level of high automation. 6. Tagung Fahrerassistenzsysteme. *Der Weg zum Autom. Fahren*. (2013)
21. Aeberhard, M., et al.: Experience, results and lessons learned from automated driving on Germany's highways. *IEEE Intell. Transp. Syst. Mag.* **7**(1), 42–57 (2015)
22. Schmidt, A., Dey, A.K., Kun, A.L., Spießl, W.: Automotive user interfaces: human computer interaction in the car. In: *CHI 2010 Extended Abstracts on Human Factors in Computing Systems (CHI EA 2010)*, pp. 3177–3180 (2010)
23. Okuda, R., Kajiwara, Y., Terashima, K.: A survey of technical trend of ADAS and autonomous driving. In: *Proceedings of Technical Program - 2014 International Symposium on VLSI Technology, Systems and Application, VLSI-TSA 2014* (2014)
24. Naujoks, F., Purucker, C., Neukum, A.: Secondary task engagement and vehicle automation - comparing the effects of different automation levels in an on-road experiment. *Transp. Res. Part F Traffic Psychol. Behav.* **38**, 67–82 (2016)
25. Tesla: Tesla. <https://www.tesla.com/modelx/design>. Accessed 30 Oct 2016
26. Boesler, M.: The 27 Best Selling Vehicles in America, *Business Insider* (2012)
27. NHTSA, U.S. Department of Transportation Releases Policy on Automated Vehicle Development, National Highway Traffic Safety Administration (NHTSA) (2013). <http://www.nhtsa.gov/About+NHTSA/Press+Releases/U.S.+Department+of+Transportation+Releases+Policy+on+Automated+Vehicle+Development>. Accessed 30 May 2013
28. S. International, Core Principles (2016). <http://saannualreport.org/2015/overview/mission-vision-statements/>

29. Summers, S., Prasad, A., Hollowell, W.T.: NHTSA's vehicle compatibility research program, SAE Technical Paper (1999)
30. NHTSA, NHTSA Statutory Authorities. <https://www.nhtsa.gov/Laws-&-Regulations/NHTSA-Statutory-Authorities>. Accessed 30 Oct 2016
31. SAE International, Automated Driving
32. Boston Consulting Group, A Roadmap to Safer Driving Through Advanced Driver Assistance Systems Safer Driving (2015)
33. Tigadi, A., Gujanatti, R., Gonchi, A.: Advanced driver assistance systems. *Int. J. Eng. Res. Gen. Sci.* **4**(3), 151–158 (2016)
34. Naujoks, F., Mai, C., Neukum, A.: The effect of urgency of take-over requests during highly automated driving under distraction conditions. In: Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics, AHFE, pp. 2099–2106, July 2014
35. Naujoks, F., Purucker, C., Neukum, A., Wolter, S., Steiger, R.: Controllability of partially automated driving functions - Does it matter whether drivers are allowed to take their hands off the steering wheel? *Transp. Res. Part F Traffic Psychol. Behav.* **35**, 185–198 (2015)
36. Morignot, P., Rastelli, J.P., Nashashibi, F.: Arbitration for balancing control between the driver and ADAS systems in an automated vehicle: survey and approach. In: Proceedings of the IEEE Intelligent Vehicles Symposium, pp. 575–580 (2014)
37. Martens, M.H., Van Den Beukel, A.P.: The road to automated driving: dual mode and human factors considerations. In: Proceedings of the IEEE Conference on Intelligent Transportation Systems, ITSC, pp. 2262–2267 (2013)
38. Pink, O., Becker, J., Kammel, S.: Automated driving on public roads: experiences in real traffic. *IT - Inf. Technol.* **57**(4), 223–230 (2015)
39. Lee, K.J., Joo, Y.K., Nass, C.: Partially intelligent automobiles and driving experience at the moment of system transition. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI 2014, pp. 3631–3634 (2014)
40. Stanton, N.A., Young, M.S.: Driver behaviour with adaptive cruise control. *Ergonomics* **48** (10), 1294–1313 (2005)
41. Sheridan, T.B., Parasuraman, R.: Human-automation interaction. *Rev. Hum. Factors Ergon.* **1**(1), 89–129 (2005)
42. Lin, P.: Why Ethics Matters for Autonomous Cars. In: *Autonomes Fahren*, pp. 70–85 (2015)
43. Hevelke, A., Nida-Rümelin, J.: Responsibility for crashes of autonomous vehicles: an ethical analysis. *Sci. Eng. Ethics* **21**(3), 619–630 (2015)
44. Goodall, N.: Ethical decision making during automated vehicle crashes. *Transp. Res. Rec. J. Transp. Res. Board* **2424**, 58–65 (2014)
45. Vasic, M., Billard, A.: Safety issues in human-robot interactions. In: 2013 IEEE International Conference on Robotics and Automation (ICRA), pp. 197–204 (2013)
46. Alvarez, I., Alnizami, H., Dunbar, J., Jackson, F., Gilbert, J.E.: Help on the road: effects of vehicle manual consultation in driving performance across modalities. *Int. J. Hum. Comput. Stud.* **73**, 19–29 (2015)
47. Kujala, T., Salvucci, D.D.: Modeling visual sampling on in-car displays: the challenge of predicting safety-critical lapses of control. *Int. J. Hum. Comput. Stud.* **79**, 66–78 (2015)
48. N. Highway Traffic Safety Administration and D. of Transportation, Research Note: Distracted Driving 2014 (2014)
49. 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. *Cogn. Technol. Work* **14**(1), 3–18 (2012)
50. Horne, J., Reyner, L.: Vehicle accidents related to sleep: a review. *Occup. Environ. Med.* **56**, 289–294 (1999)

51. Shchetko, N.: Laser eyes pose price hurdle for driverless cars size and price of lidar technology will have to fall for autonomous vehicles to thrive. *Wall Str. J.* (2014)
52. Scotchmer, S., Green, J.: Novelty and disclosure in patent law. *RAND J. Econ.* **21**(1), 131–146 (1990)
53. Gans, J.S., Murray, F.: Funding Scientific Knowledge Selection, Disclosure, and the Public-Private Portfolio, March 2012
54. A.A. News, Witnesses to aftermath of deadly Tesla say AutoPilot continued to drive car for hundreds of yards (2016)
55. T. Drive, Another Person Has Reportedly Died Due to a Tesla Autopilot Failure (2016)
56. T. Drive, Watch Yet Another Tesla Model S Crash While on Autopilot (2016). <http://www.thedrive.com/news/3679/watch-yet-another-tesla-model-s-crash-while-on-autopilot>
57. T. Drive, Tesla Driver Claims Autopilot Responsible for This Dashcam Crash (2016). <http://www.thedrive.com/news/4731/tesla-driver-claims-autopilot-responsible-for-this-dashcam-crash>
58. Lee, J.D., See, K.A., City, I.: Trust in automation: designing for appropriate reliance. *Hum. Factors J. Hum. Factors Ergon. Soc.* **46**(1), 50–80 (2004)
59. A.T. Association, Report, Trends & Statistics (2016). http://www.trucking.org/News_and_Information_Reports_Industry_Data.aspx
60. Forrest, A., Konca, M.: *Autonomous Cars and Society* (2007)
61. U.S.D. of Labor, Occupational Outlook Handbook: Bus Drivers (2014). <http://www.bls.gov/oooh/transportation-and-material-moving/bus-drivers.htm>
62. U.S.D. of Labor, Occupational Outlook Handbook: Taxi Drivers and Chauffeurs (2014). <http://www.bls.gov/oooh/transportation-and-material-moving/taxi-drivers-and-chauffeurs.htm>
63. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. *Int. J. Hum. Comput. Stud.* **58**(6), 697–718 (2003)
64. Thill, S., Nilsson, M., Riveiro, M.: Perceived intelligence as a factor in (semi-) autonomous vehicle UX. In: CHI 2010 Extended Abstracts on Human Factors in Computing Systems (CHI EA 2010) (2015)
65. Muir, B.M.: Trust between humans and machines, and the design of decision aids. *Int. J. Man Mach. Stud.* **27**, 527–539 (1987)
66. Helldin, T., Falkman, G., Riveiro, M., Davidsson, S.: Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In: Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI 2013, vol. 5, pp. 210–217 (2013)
67. Glancy, D.J.: Privacy in autonomous vehicles. *Santa Clara Law Rev.* **1**(4), 77 (2012)
68. Boeglin, J.: The cost of self-driving cars: reconciling freedom and privacy with tort liability in autonomous vehicle regulation. *Yale J. Law Technol.* **17**(171), 30 (2015)
69. Greenberg, A.: Hackers remotely kill a jeep on the highway—with me in it, *Wired* (2015)
70. Kirkpatrick, K.: The moral challenges of driverless cars. *Commun. ACM* **58**(8), 19–20 (2015)
71. Degani, A., Heymann, M.: Formal verification of human-automation interaction. *Hum. Factors* **44**(1), 28–43 (2002)
72. Sarter, N.B., Woods, D.: Pilot interaction with cockpit automation - operational experiences with the flight management system. *Int. J. Aviat. Psychol.* **2**(4), 37–41 (1992)
73. Pausie, A.: A method to assess the driver mental workload: the driving activity load index (DALI). *IET Intell. Transp. Syst.* **2**(4), 315–322 (2008)
74. Hughes, D., Dornheim, M.A.: Accidents direct focus on cockpit automation. *Aviat. Space Technol.* **142**, 52–54 (1995)

75. Flemisch, F., Kelsch, J., Löper, C., Schieben, A., Schindler, J.: Automation spectrum, inner/outer compatibility and other potentially useful human factors concepts for assistance and automation. *Hum. Factors Assist. Autom.* **2008**, 1–16 (2008)
76. Billings, C.: *Aviation Automation: The Search for a Human-Centered Approach*. Lawrence Erlbaum Associates, Mahwah (1997)
77. Navaro, J., Mars, F., Young, M.S.: Lateral control assistance in car driving: classification, review and future prospects. *IET Intell. Transp. Syst.* **5**(3), 207 (2011)
78. Harris, D., Harris, F.J.: Evaluating the transfer of technology between application domains: a critical evaluation of the human component in the system. *Technol. Soc.* **26**(4), 551–565 (2004)
79. Carsten, O.M.J., Tate, F.N.: Intelligent speed adaptation: accident savings and cost-benefit analysis. *Accid. Anal. Prev.* **37**(3), 407–416 (2005)
80. Nilsson, L.: Safety effects of adaptive cruise controls in critical traffic situations. In: ‘Steps Forward’. Proceedings of the Second World Congress on Intelligent Transport Systems 1995, vol. 3, Yokohama, pp. 1254–1259 (1995)
81. Martin, E.: From Zero Hours to Airline Pilot: How Long will it Take? <https://www.pea.com/blog/posts/zero-hours-airline-pilot-long-will-take/>
82. Williams, A.F., Weinberg, K., Fields, M., Ferguson, S.A.: Current requirements for getting a drivers license in the United States. *J. Saf. Res.* **27**(2), 93–101 (1996)
83. Goodrich, M.A., Schultz, A.C.: Human-robot interaction: a survey. *Found. Trends® Hum. Comput. Interact.* **1**(3), 203–275 (2007)
84. Merat, N., Lee, J.D.: Preface to the special section on human factors and automation in vehicles: designing highly automated vehicles with the driver in mind. *Hum. Factors J. Hum. Factors Ergon. Soc.* **54**(5), 681–686 (2012)

How Do Hybrid Electric Vehicle Drivers Acquire Ecodriving Strategy Knowledge?

Thomas Franke¹✉, Matthias G. Arend², and Neville A. Stanton³

¹ Institute for Multimedia and Interactive Systems, Engineering Psychology and Cognitive, Ergonomics, Universität zu Lübeck, Lübeck, Germany
franke@imis.uni-luebeck.de

² Department of Psychology, Cognitive and Engineering Psychology, Chemnitz University of Technology, Chemnitz, Germany
matthias-georg.arend@s2013.tu-chemnitz.de

³ Transportation Research Group, Faculty of Engineering and the Environment, University of Southampton, Southampton, UK
n.stanton@soton.ac.uk

Abstract. Hybrid electric vehicles (HEVs) have the potential to accomplish high energy efficiency (i.e., low fuel consumption) given that drivers apply effective ecodriving control strategies (i.e., ecodriving behavior). However, HEVs have a relatively complex powertrain and therefore require a considerable knowledge acquisition process to enable optimal ecodriving behavior. The objective of the present research was to examine the acquisition of ecodriving strategy knowledge in HEV drivers who are successful in achieving a relatively high energy efficiency. To this end, we recruited 39 HEV drivers with above-average fuel efficiencies and collected interview data on the ecodriving strategy acquisition process. Drivers reported the acquisition of different types of knowledge as important for ecodriving, namely specific strategy knowledge and general technical system knowledge. They acquired this knowledge both with system-interaction (e.g., actively testing specific strategies, continuous monitoring of energy consumption) and without system-interaction (e.g., internet forums, consulting experts). This learning process took drivers on average 6.4 months or 10062 km. The results show the high diversity of the means that HEV drivers use to develop their ecodriving knowledge and the considerable time it takes HEV drivers to develop their ecodriving strategies.

Keywords: Hybrid electric vehicles · Ecodriving · Strategy knowledge · Learning process · Driving behavior

1 Introduction

Road transport is one of the major factors contributing to global CO₂ emissions caused by energy infrastructure [7]. Electric vehicles constitute a promising technology that has the potential to reduce road transport CO₂ emissions considerably (e.g., [4]). Indeed, electrification of road transport is currently one of the largest technological transformations in the field of sustainable development. However, to make this transformation a success the user factor has to be taken into account in system design.

That is, each technical system designed to foster sustainability can be characterized by its technical sustainability potential (e.g., maximum possible energy efficiency under realistic usage conditions). Yet, in the end it often depends on user behavior (e.g., an optimal interaction with the system features designed to optimize energy efficiency) whether this potential can actually be realized in everyday usage of the system.

Ecodriving is the term that is commonly used to describe all those driving behaviors that are performed to increase real-world energy efficiency of a road vehicle (e.g. [3, 14, 25, 26, 30]). In electric vehicles, this influence of driver behavior on energy efficiency has been discussed as particularly relevant as research suggests that drivers have a high impact on energy consumption in electric vehicles compared to conventional combustion vehicles [21, 22, 31]. Further, electric powertrains have specific energy dynamics (i.e., powertrain efficiency characteristics, [16]) that make behavioral adaptation necessary to achieve a high energy efficiency. Hence, drivers have to understand these energy efficiency characteristics of the powertrain to adapt their behavior accordingly.

Hybrid electric vehicles (HEVs) represent the most complex electric powertrain and at the same time constitute an increasingly widespread type of electric vehicle [1] which makes it particularly relevant to address potential human factors issues associated with their usage. The complexity of this powertrain mainly results out of the combination of an electric motor and a combustion engine that interact at a fast rate as well as a regenerative braking system that creates a highly dynamic bidirectional energy flow. Those powertrain characteristics can be presumed to require behavioral adaptation [2, 6]. Thus, HEV drivers are likely in particular need of a sufficiently precise understanding of the energy dynamics to perform effective ecodriving strategies in HEV usage (see also [9]). Hence, there is a considerable risk that a substantial part of the sustainability potential of HEVs is lost due to a lack of drivers' understanding of effective ecodriving strategies.

A key challenge from the perspective of green ergonomics [13, 28] therefore is to advance understanding of user-energy interaction in HEVs and how drivers can be best supported in the acquisition of correct mental models and effective ecodriving strategies. One step in this research agenda is to gain a comprehensive understanding of drivers' adaptation to the system and the development of ecodriving strategy knowledge.

The objective of the present research was to examine the acquisition of ecodriving strategy knowledge in HEV drivers. To specifically gain an understanding of successful learning strategies we concentrated on drivers who are particularly successful in optimizing the energy efficiency of their HEVs. To this end, we recruited 39 HEV drivers who achieved an above-average fuel efficiency with their HEV under everyday conditions compared to a population of HEV drivers who regularly logged their fuel consumption data (i.e. all could be assumed to be at least somewhat interested in energy efficiency).

2 Background

Several challenges have to be overcome until a driver can achieve a high energy efficiency with a particular vehicle. From a motivation and volition perspective of goal setting, goal striving and action regulation these can mainly be broken down into three steps. First, drivers need to cognitively represent the goal of ecodriving in an accessible way and evaluate it positively and achievable (i.e., set ecodriving as a salient goal for establishing a basic level of ecodriving motivation). Second, drivers need to pursue the ecodriving goal with high priority in comparison to other attractive (i.e. competing) goals such as time-efficiency, safety, comfort or driving pleasure ([8]; i.e., volitional processes, [18]). However, the largest challenge for drivers that potentially demands the most cognitive resources is the continuous control of goal-directed behavior to increase energy efficiency. Interestingly, while considerable research so far has focused on increasing and sustaining ecodriving motivation (e.g. [17, 26]) there is less research that focusses on the issues of cognitive control of ecodriving behavior on a more microscopic level (for exceptions see e.g., [2, 9, 19]).

Previously, we have suggested that ecodriving efforts can be conceptualized as a control theoretic feedback loop [9] parallel to the increasingly widespread notions in driver behavior models that examine on other facets of driving behavior, typically safe/unsafe driving, from a control theory perspective [11, 12, 27]. With regard to the regulation of ecodriving behavior it can be conceptualized that the control process starts with the perception of relevant environmental variables (i.e., system state and state of environment) to enable the identification of applicable ecodriving strategies in a given situation. Further, drivers should select and finally implement a strategy that is perceived as particularly energy-efficient in a given situation (and fits to other motives/goals such as safety or comfort).

A core component of this framework is the strategy knowledge base that is expected to comprise the drivers' repertoire of strategies (i.e., which ecodriving strategies exist), as well as their subjective conceptualizations of energy efficiency (i.e., which ecodriving strategies are effective and why). A relevant task for advancing understanding of the ecodriving control process is thus, to examine this knowledge base and how it is established (i.e., how ecodriving strategy knowledge is developed).

First research in this area has examined drivers' knowledge and mental models of energy efficiency in driving conventional vehicles [20, 23] showing, for example, that ecodriving knowledge in the general population is rather low [20] yet when asked to drive energy efficient, drivers change their driving behavior (compared to normal driving behavior; [23]). Furthermore, first research in the context adaption to electric vehicle driving has demonstrated that ecodriving knowledge is dynamic and develops with practical driving experience [22, 24], or with specific supporting ecodriving feedback [14, 15]. However, research addressing the question of how exactly drivers acquire this ecodriving knowledge is still lacking, in particular with regard to user-interaction with new and complex powertrains like HEVs.

3 Method

3.1 Participants

We focused recruitment on HEV drivers of the Toyota Prius (2nd gen, 3rd gen, and Prius c [in Germany sold as Yaris Hybrid]), being the most sold (see e.g., [29]) and most prototypical HEV model. To enable recruitment of drivers who achieved a specific fuel efficiency with their HEV we used the database on www.spritmonitor.de.

From the almost 1500 Prius drivers in the database we invited drivers who (a) had an average fuel efficiency above the fleet-average of the vehicle model, (b) were from Germany, Austria, or Switzerland, and (c) had logged their fuel efficiency within the last 3 months. We avoided drivers who appeared to log fuel efficiency inconsistently, and sought to sample drivers across a range of above-average fuel efficiencies (i.e., from “just above average” to “top of the list”). Ethical approval was sought from and granted by the University of Southampton’s Ethics and Research Governance committee (reference number 17071).

Participants in the resulting sample ($N = 39$) had an average age of $M = 45$ years ($SD = 10$) and an average HEV driving experience of $M = 74079$ km ($SD = 64513$), 92% were male, and 56% had a university degree.

3.2 Procedure

To collect the required, data telephone interviews (including questionnaire sections) were conducted ($M_{\text{duration}} = 48$ min, $SD = 8$). Participants received the interview guideline before the interview and could therefore refer to the documentation as the interviewer went through the questions. The interviewer’s experience with HEV driving (>6 years) facilitated the interview process.

After introducing the study and gaining informed consent, the interview had the following parts: (P1) ecodriving motivation, (P2) ecodriving strategies, (P3) questions on ecodriving strategy development, false beliefs (i.e. false mental models), and user suggestions for ecodriving support systems (P4) questionnaire parts to assess socio-demographic and experience-related variables. The interview was audio-recorded and transcribed. The present paper focuses on the part of section (P3) of the interview that deals with strategy acquisition (development of ecodriving strategies). Results regarding the other parts of the interview and further details on the methodology have been published in [9, 10].

3.3 Qualitative Data Analysis

We based our qualitative data analysis on thematic analysis [5]. After each interview, the interviewer and the scribe (first and second author) discussed insights and first ideas for possible codes. After familiarization with the data, the initial coding phase led to a list of codes that was relevant to our research question (consequently, only statements referring to the acquisition of knowledge were provided with codes). Afterwards, the coding system was reviewed and discussed, and initial ideas for themes (i.e., thematic

clusters) were revised and refined based on the thematic proximity of codes. As only a relatively low level of abstraction of statements was targeted, this phase was less complex than for other topics in psychology (i.e., semantic rather than latent level analysis; [5]). In the final phase, we again went through all transcripts and coded participants' statements with regard to the developed coding systems (i.e., clusters and sub-clusters). Within this phase some final revisions and refinements of the coding system were performed. Hence, all statements of the participants that were relevant to the respective research question were grouped into clusters and sub-clusters. Clusters group similar statements of different participants (i.e., an overarching theme that is addressed by several participants).

4 Results

All percentage values given in the following sections refer to the share of drivers from the whole sample ($N = 39$). Note that percentages must not add up to 100% because one driver can use more than one mean (i.e., no exclusive coding was performed).

To address our research question, how HEV drivers acquire ecodriving strategy knowledge, we posed the following question to drivers: "How did you developed your strategies over time?". This question directly followed the extensive interview section where drivers elaborated on (a) the ecodriving strategies they used in four prototypic situations (driving on the autobahn, city driving, relatively straight rural road with flat terrain, mountainous and winding rural road) as well as (b) their conceptualization about why these strategies were effective in increasing energy efficiency (for results see [9]). Hence, it can be assumed that this elaboration of the ecodriving strategy knowledge base also made the memories about how the ecodriving strategies developed over time more accessible in memory and, therefore, helped drivers to sketch a more comprehensive and precise picture about their ecodriving strategy acquisition.

With regard to results, first, drivers' answers to the interview question were coded as the acquisition of two knowledge types: (a) strategy knowledge (95%) and (b) relevant technical system knowledge (41%). Second, another key distinction that resulted from the data was, whether strategy or technical system knowledge were reported to having been acquired *with interaction* with the HEV (69%) or *without interaction* with the HEV (87%). Hence, this pattern can be structured in a fourfold table. Consequently, a correspondingly structured overview of the main means that drivers used to acquire their ecodriving strategies is depicted in Table 1.

In the following sections, a more detailed view on how drivers acquired strategy and technical system knowledge will be presented. We refer to the fourfold table above for structuring the results. For each of the four categories, descriptions of those categories, sub-topics (i.e., acquisition means) and example statements are given.

4.1 The Acquisition of Strategy Knowledge

In the present study, strategy knowledge can be defined as the knowledge about driving behaviors that are perceived efficient. Obviously, this knowledge does not refer to

Table 1. General overview of acquisition means used to acquire ecodriving knowledge

		Knowledge type	
		Strategy knowledge	Technical system knowledge
Acquisition Means	With system-interaction	Testing Monitoring Incidental learning System trains driver	Testing
	Without system-interaction	Internet forums Expert survey Additional literature Videos Former knowledge	Internet forums Additional literature Videos

Note. This table gives a first overview of the results; more detailed results are presented in the text below.

strategies which undoubtedly have a positive effect on real-world fuel efficiency, but to knowledge about strategies that is deemed efficient by the drivers. As stated above, strategy knowledge has been acquired with and without HEV system-interaction (see Table 1). We consequently present the results in the order of cells of this fourfold table.

Acquisition of Strategy Knowledge with Interaction with the HEV. Many drivers (64.1%) reported that they have acquired strategy knowledge during their trips with the HEV. A crucial role in this respect played the active testing of different strategies which was reported to be used by more than half of the drivers (54%). To infer the efficiency of the strategies tested, the drivers reported to monitor different kinds of system feedback, of which fuel consumption was the most prominent (52%). Yet, there were some differences in regard of which specific fuel consumption criterion drivers monitored: fuel consumption per route (e.g., on the daily route; 28%) fuel consumption in general (e.g., on the basis of tank fillings recorded; 23%) and fuel consumption within very short time periods (e.g., instantaneous fuel consumption; 15%). In this respect, drivers used different kinds of system feedback (e.g., the fuel consumption history display that is provided by the HEV models in the present study) or other helpful applications (e.g., spritmonitor.de website to log the fuel consumption data):

“I drive the same route each day. [...] Through the instantaneous fuel consumption as it is currently displayed in the vehicle, I get information and obviously also by the curve [refers to fuel consumption history display]. The higher the resolution, the better. Next, I have been using spritmonitor.de for many, many years and through this I was obviously able to monitor even better and, in terms of fuel ... from refueling stop to refueling stop examine, how it develops.”(P03)

Furthermore, to test the effectiveness of the specific ecodriving strategy pulse and glide (drivers repeatedly accelerate in a pulse-phase to a target speed, subsequently decrease this speed in the glide phase, see [9]), even specific driving environments and conditions were used to control for further influences:

“And then I obviously conducted a test, at night, with low traffic, the influence of driving only with pulse and glide between 30 and 50 km/h. Lo and behold, I get down to 3.2 l/100 km.” (P05)

Beyond fuel consumption, drivers also monitored other kinds of system feedback to acquire strategy knowledge: further feedback provided by the hybrid system (e.g., depiction of the energy flows; 15%) and by additional apps and tools (10%) and sounds (3%). How the feedback by the hybrid system affects ecodriving strategies (e.g. intensity of regenerative braking), is illustrated by the following statement:

“There are the displays in the car which really help me, these are the energy bar display where one can see where the vehicle saves energy, how it also saves energy in braking. When I brake strongly the system recharges less, when I brake less intense, one can see that the energy recovery is a little bit bigger [...]” (P19)

In contrast to the active testing of different strategies, some drivers also mentioned kind of a passive strategy knowledge acquisition process: Drivers reported that strategy effectiveness is better understood through incidental learning by interacting with the HEV (*“it is learning by doing”*, [P09]; 10.3%), or that the system is educating the driver (*“I’ve got the feeling, that the car trains the driver in driving efficiently. Simply because of its functionality”*[P27]; 7.7%).

Acquisition of Strategy Knowledge without Interaction with the HEV. When acquiring strategy knowledge without system-interaction, drivers reported to rely on internet forums (64%), their former knowledge (39%), ‘expert’ surveys (asking experts, e.g. experienced drivers, about their knowledge; 8%), reading additional literature (e.g., the user manual 8%) or videos (5%).

From the large percentage of drivers using internet forums to acquire strategy knowledge, a considerable share (18%) already informed themselves prior to the purchase. Moreover, for some drivers it could be categorized which specific strategies they had adopted from those forums. Those strategies were pulse and glide (8%), a specific way of accelerating from standstill (5%), using (or rather not using) the B-mode (engine break active, cf. Franke et al., 2016), utilizing electric driving, the adaptive cruise control and neutral coasting in the neutral mode (each 3%). This process of acquiring knowledge through the internet and applying this knowledge to real word driving is reflected in the following statement:

“Then I have learned pulse and glide, in the third phase. I found this strategy not in the manual of Toyota but on the internet and I have to say, this [pulse and glide] poses, to me, the most powerful tool among those strategies I have learned, based on the driving technique. With which one can really influence the fuel consumption most.” (P31)

Some drivers also reported that former knowledge played a crucial role in developing HEV strategy knowledge, particularly ecodriving strategies they have already acquired with their former vehicle (26%) and physics knowledge (8%). The former primarily refers to some general ecodriving knowledge applicable to most vehicles, as stated by one driver:

“I guess that I have already developed most of the strategies with my conventional vehicle” (P16)

The latter, physics knowledge, depicts a general influence on the strategy knowledge in a way that strategy effectiveness can be deduced from it, as reflected in the statement of the following driver:

“I was attentive during the physics lessons and I know what consumes energy and what doesn’t. So, I know that I need the most energy for accelerating and for constantly gliding at one speed, I only have to overpower the friction and air resistance and I consume relatively few [fuel].” (P12)

Beyond those factors already mentioned, several drivers (26%) perceived personal attributes as important determinants of their acquisition of strategy knowledge. Below them was the fun factor of driving energy efficient (10%), as stated by another driver:

“One more often remembers to drive in that way [energy efficient] because it is fun. So, one is tempted to try ... ‘maybe I can reduce my fuel consumption for another 0.2 l/100 km’” (P32)

Other attributes mentioned were trust (i.e., knowledge that the algorithms will control the vehicle most energy efficient; 3%) or the drivers interest in the technology (3%).

4.2 The Acquisition of Technical System Knowledge

The acquisition of technical system knowledge (concerning the HEV system) is a construct that was subject in former research and has proven an important determinant of HEV ecodriving success [2, 9]. Consequently, as drivers did not exclusively rely on reporting the acquisition of strategy knowledge, we also analyzed their reports on the acquisition of this knowledge type. As for strategy knowledge, reports for technical system knowledge referred to acquisition with and without interaction with the system. Thus, results will again be presented in this order.

Acquisition of Technical System Knowledge with Interaction with the HEV.

Several drivers (15%) reported that they have acquired strategy knowledge through the interaction with the system. For example, the way system dynamics (e.g., the state of charge) influence how the driving energy is supplied might be understood through interacting with the HEV system:

“I now know how far I may press the accelerator pedal before the combustion engine will turn on, furthermore I also know now that this depends on the state of charge of the battery. It is always less optimal the emptier the battery becomes. And it is more likely that the combustion engine will also turn on. This is taught by the car.” (P28)

Acquisition of Technical System Knowledge without Interaction with the HEV.

Finally, a total of 13 drivers (33%) reported that they have gained technical system knowledge through sources like internet forums (28%), professional literature (3%) and videos (of the hybrid system; 3%).

“Yes, I am an active member of the Priusfreunde-forum. [...] They also provide the Prius-Wiki, where many technical aspects are explained. This has helped my understanding a lot.”(P04)

4.3 Duration of Learning Process

Finally, in addition to the main interview question of strategy acquisition, we asked drivers the sub-question how long it took them to acquire their ecodriving strategies (with regard to time and total distance driven). This information could finally be derived from $N = 38$ participants, because one participant did not provide a clear numerical value. Further three drivers were not aware of a substantial strategy learning process during the time of their HEV usage, meaning they perceived that they had already acquired the ecodriving knowledge they deemed necessary prior to purchasing their HEV (e.g., “Prior to purchasing the car [...] I had concerned myself certainly for 1.5 years very extensively with the topic of hybrids, thus, purely hypothetical.” [P20] or “I was attentive during the physics lessons.” [P12]). Within the sample of the remaining $N = 35$ drivers, the average time it took them to develop their ecodriving strategies, based on the drivers’ estimates, was $M = 6.4$ months ($SD = 7.8$, 25th percentile = 2, 75th percentile = 12 months). Moreover, drivers’ duration estimates in terms of total distance driven were $M = 10062$ km ($SD = 8785$, 25th percentile = 4000, 75th percentile = 17000 km).

5 Discussion

How and where do HEV drivers learn to optimize the energy efficiency of their vehicles? Is it enough to hear or read some basic rules to become a successful eco-driver, or which learning strategies do HEV drivers really use? This was the question of the present research. The results showed that drivers used a diverse set of means to develop their ecodriving control strategies and that this learning process took them considerable time.

The successful ecodrivers we recruited for our interview study (i.e., drivers with above-average fuel efficiency compared to a sample of HEV drivers who could all be assumed to have a basic interest in their fuel consumption) did not only report the learning of specific rules or strategies to be important for acquiring ecodriving skills, yet also stated that technical system knowledge (e.g., understanding of the technical interplay of the powertrain components) played an important role. Hence, it appears that, to support development of ecodriving skills in novice drivers of HEVs, these drivers should not only be provided with ecodriving tips focused on specific behaviors (i.e., “this behavioral strategy is effective to optimize energy efficiency”). Instead, drivers should also be provided with the necessary background knowledge to understand why these behavioral strategies work (e.g., “there are two states where the combustion engine in your HEV is most efficient and that is why you should target these states which you can do with this strategy”, or “air resistance increases exponentially with speed, hence reducing speed on the motorway is very effective”). This could also, for example, make drivers more flexible in applying certain strategies and derive new ones.

Moreover, ecodriving strategy knowledge is developed both while driving as well as without direct system interaction. In particular, the time-consuming active testing and monitoring of certain strategies shows that identifying the optimal ecodriving strategies is not that simple. Variations in environmental conditions (e.g., presence and

behavior of other vehicles) make it complex for drivers to clearly trace the success of the application of certain ecodriving strategies. Here, there is a considerable potential of ecodriving support systems such as systems that help to objectively quantify environmental characteristics with relevance for energy consumption (e.g., traffic flow, road conditions, terrain) and depict these along with energy consumption data (e.g., in an energy consumption history display) to aid drivers in disentangling effects of their behavior versus environmental factors on fuel consumptions.

Further, the results of the present study indicated that control and monitoring processes of ecodriving strategy effects are implemented at different time scales, that is, drivers use different aggregations (i.e., reference periods) to check the success of their ecodriving efforts such as the consumption averaged over one trip or one filling of the tank (i.e., one refueling cycle), down to consumption in the last minute or second. These different levels of control have to be considered in system design as well as in theorizing of ecodriving behavior. Future research could be concerned with which monitoring reference periods provide adequate feedback for ecodriving success.

Finally, the process of developing ecodriving knowledge in HEV driving (at least the time that successful ecodrivers need to acquire their level of knowledge) takes considerable time. This echoes the notion that the powertrain of HEVs is complex and energy dynamics are difficult to understand for drivers. Moreover, one can assume that the development of ecodriving strategies is not just a matter of knowing a certain strategy but also acquiring the knowledge how to actually implement this strategy in different situations. Hence, each ecodriving support system that is focused on supporting strategy development should not only be focused on teaching rules and strategies, but also continuously guide drivers' in their efforts to implement these strategies in different driving situations.

All in all, the present study represents a first step to advance understanding of ecodriving strategy development in HEV drivers. Further quantitative studies will be needed to assess the processes and patterns discussed in the present contribution with greater precision and provide further insights into the questions of how achieving optimal energy efficiency and reaching the technical sustainability potential of a system can be facilitated as much as possible for users.

Acknowledgments. This research was partly supported by a DAAD grant to the first author and an ERASMUS + grant to the second author. We gratefully thank Prof. Dr. Josef Krems for providing parts of the research infrastructure.

References

1. Al-Alawi, B.M., Bradley, T.H.: Review of hybrid, plug-in hybrid, and electric vehicle market modeling studies. *Renewable Sustain. Energy Rev.* **21**, 190–203 (2013). doi:[10.1016/j.rser.2012.12.048](https://doi.org/10.1016/j.rser.2012.12.048)
2. Arend, M.G., Franke, T.: The role of interaction patterns with hybrid electric vehicle eco-features for drivers' ecodriving performance. *Hum. Factors* (2016). doi:[10.1177/0018720816670819](https://doi.org/10.1177/0018720816670819)

3. Barkenbus, J.N.: Eco-driving: An overlooked climate change initiative. *Energy Policy* **38**, 762–769 (2010). doi:[10.1016/j.enpol.2009.10.021](https://doi.org/10.1016/j.enpol.2009.10.021)
4. Bitsche, O., Gutmann, G.: Systems for hybrid cars. *J. Power Sources* **127**, 8–15 (2004). doi:[10.1016/j.jpowsour.2003.09.003](https://doi.org/10.1016/j.jpowsour.2003.09.003)
5. Braun, V., Clarke, V.: Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**, 77–101 (2006). doi:[10.1191/1478088706qp063oa](https://doi.org/10.1191/1478088706qp063oa)
6. Cocron, P., Bühler, F., Neumann, I., Franke, T., Krems, J.F., Schwalm, M., Keinath, A.: Methods of evaluating electric vehicles from a user's perspective—the MINI E field trial in Berlin. *IET Intel. Transport Syst.* **5**, 127–133 (2011). doi:[10.1049/iet-its.2010.0126](https://doi.org/10.1049/iet-its.2010.0126)
7. Davis, S.J., Caldeira, K., Matthews, H.D.: Future CO₂ emissions and climate change from existing energy infrastructure. *Science* **329**, 1330–1333 (2010). doi:[10.1126/science.1188566](https://doi.org/10.1126/science.1188566)
8. Dogan, E., Steg, L., Delhomme, P.: The influence of multiple goals on driving behavior: the case of safety, time saving, and fuel saving. *Accid. Anal. Prev.* **43**, 1635–1643 (2011). doi:[10.1016/j.aap.2011.03.002](https://doi.org/10.1016/j.aap.2011.03.002)
9. Franke, T., Arend, M.G., McIlroy, R.C., Stanton, N.A.: Ecodriving in hybrid electric vehicles—exploring challenges for user-energy interaction. *Appl. Ergon.* **55**, 33–45 (2016). doi:[10.1016/j.apergo.2016.01.007](https://doi.org/10.1016/j.apergo.2016.01.007)
10. Franke, T., Arend, M.G., McIlroy, R.C., Stanton, N.A.: What drives ecodriving? hybrid electric vehicle drivers' goals and motivations to perform energy efficient driving behaviors. In Stanton, N.A., Landry, S., Di Bucchianico, G., Vallicelli, A. (eds.) *Advances in Human Aspects of Transportation*. AISC, vol. 484, pp. 451–461. Springer, London (2016b). doi:[10.1007/978-3-319-41682-3_38](https://doi.org/10.1007/978-3-319-41682-3_38)
11. Fuller, R.: Motivational determinants of control in driving task. In: Cacciabue, P.C. (ed.) *Modelling Driver Behaviour in Automotive Environments: Critical Issues in Driver Interactions with Intelligent Transport Systems*, pp. 165–188. Springer, London (2007). doi:[10.1007/978-1-84628-618-6_10](https://doi.org/10.1007/978-1-84628-618-6_10)
12. Fuller, R.: Driver control theory: from task difficulty homeostasis to risk allostasis. In: *Handbook of Traffic Psychology*, pp. 208–232. Elsevier, Amsterdam (2011). doi:[10.1016/B978-0-12-381984-0.10002-5](https://doi.org/10.1016/B978-0-12-381984-0.10002-5)
13. Hanson, M.A.: Green ergonomics: challenges and opportunities. *Ergonomics* **56**, 399–408 (2013). doi:[10.1080/00140139.2012.751457](https://doi.org/10.1080/00140139.2012.751457)
14. Jamson, S.L., Hibberd, D.L., Jamson, A.H.: Drivers' ability to learn eco-driving skills; effects on fuel efficient and safe driving behaviour. *Transp. Res. Part C: Emerg. Technol.* **58**, 657–668 (2015). doi:[10.1016/j.trc.2015.02.004](https://doi.org/10.1016/j.trc.2015.02.004)
15. Jamson, A.H., Hibberd, D.L., Merat, N.: Interface design considerations for an in-vehicle eco-driving assistance system. *Transp. Res. Part C: Emerg. Technol.* **58**, 642–656 (2015). doi:[10.1016/j.trc.2014.12.008](https://doi.org/10.1016/j.trc.2014.12.008)
16. Kuriyama, M., Yamamoto, S., Miyatake, M.: Theoretical study on eco-driving technique for an electric vehicle with dynamic programming. In: *Proceedings of the 2010 International Conference on Electrical Machines and Systems*. pp. 2026–2030. IEEE Press, New York (2010). doi:[10.11142/jicems.2012.1.1.114](https://doi.org/10.11142/jicems.2012.1.1.114)
17. Lai, W.: The effects of eco-driving motivation, knowledge and reward intervention on fuel efficiency. *Transp. Res. Part D: Transport Environ.* **34**, 155–160 (2015). doi:[10.1016/j.trd.2014.10.003](https://doi.org/10.1016/j.trd.2014.10.003)
18. Lauper, E., Moser, S., Fischer, M., Matthies, E., Kaufmann-Hayoz, R.: Psychological predictors of eco-driving: a longitudinal study. *Transp. Res. Part F: Traffic Psychol. Behav.* **33**, 27–37 (2015). doi:[10.1016/j.trf.2015.06.005](https://doi.org/10.1016/j.trf.2015.06.005)

19. McIlroy, R.C., Stanton, N.A.: A decision ladder analysis of eco-driving: the first step towards fuel-efficient driving behaviour. *Ergonomics* **58**, 866–882 (2015). doi:[10.1080/00140139.2014.997807](https://doi.org/10.1080/00140139.2014.997807)
20. McIlroy, R.C., Stanton, N.A.: What do people know about eco-driving? *Ergonomics* (2016). doi:[10.1080/00140139.2016.1227092](https://doi.org/10.1080/00140139.2016.1227092)
21. McIlroy, R.C., Stanton, N.A., Harvey, C.: Getting drivers to do the right thing: a review of the potential for safely reducing energy consumption through design. *IET Intell. Transport Syst.* **8**, 388–397 (2014). doi:[10.1049/iet-its.2012.0190](https://doi.org/10.1049/iet-its.2012.0190)
22. Neumann, I., Franke, T., Cocron, P., Bühler, F., Krems, J.F.: Eco-driving strategies in battery electric vehicle use – how do drivers adapt over time? *IET Intell. Transport Syst.* **9**, 746–753 (2015). doi:[10.1049/iet-its.2014.0221](https://doi.org/10.1049/iet-its.2014.0221)
23. Pampel, S.M., Jamson, S.L., Hibberd, D.L., Barnard, Y.: How I reduce fuel consumption: An experimental study on mental models of eco-driving. *Transp. Res. Part C: Emerging Technol.* **58**, 669–680 (2015). doi:[10.1016/j.trc.2015.02.005](https://doi.org/10.1016/j.trc.2015.02.005)
24. Pichelmann, S., Franke, T., Krems, J.F.: The timeframe of adaptation to electric vehicle range. In: Kurosu, M. (ed.) *HCI 2013*. LNCS, vol. 8005, pp. 612–620. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-39262-7_69](https://doi.org/10.1007/978-3-642-39262-7_69)
25. Sivak, M., Schoettle, B.: Eco-driving: Strategic, tactical, and operational decisions of the driver that influence vehicle fuel economy. *Transp. Policy* **22**, 96–99 (2012). doi:[10.1016/j.tranpol.2012.05.010](https://doi.org/10.1016/j.tranpol.2012.05.010)
26. Stillwater, T., Kurani, K.S.: Drivers discuss ecodriving feedback: goal setting, framing, and anchoring motivate new behaviors. *Transp. Res. Part F: Traffic Psychol. Behav.* **19**, 85–96 (2013). doi:[10.1016/j.trf.2013.03.007](https://doi.org/10.1016/j.trf.2013.03.007)
27. Summala, H.: Towards understanding motivational and emotional factors in driver behaviour: comfort through satisficing. In: Cacciabue, P.C. (ed.) *Modelling Driver Behaviour in Automotive Environments*, pp. 189–207. Springer, London (2007). doi:[10.1007/978-1-84628-618-6_11](https://doi.org/10.1007/978-1-84628-618-6_11)
28. Thatcher, A.: Green ergonomics: definition and scope. *Ergonomics* **56**, 389–398 (2013). doi:[10.1080/00140139.2012.718371](https://doi.org/10.1080/00140139.2012.718371)
29. U.S. Department of Energy: U.S. HEV Sales by Model. <http://www.afdc.energy.gov/data/10301>
30. Young, M.S., Birrell, S.A., Stanton, N.A.: Safe driving in a green world: a review of driver performance benchmarks and technologies to support “smart” driving. *Appl. Ergon.* **42**, 533–539 (2011). doi:[10.1016/j.apergo.2010.08.012](https://doi.org/10.1016/j.apergo.2010.08.012)
31. Walsh, C., Carroll, S., Eastlake, A., Blythe, P.: Electric vehicle driving style and duty variation performance study (2010). <http://www.cenex.co.uk>

Design and Evaluation of a Mixed-Initiative Planner for Multi-vehicle Missions

Fabian Schmitt^(✉), Gunar Roth, and Axel Schulte

Institute of Flight Systems, University Bundeswehr Munich,
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
{fabian.schmitt, gunar.roth, axel.schulte}@unibw.de

Abstract. The command and control of multiple vehicles in highly dynamic scenarios by a single operator require high situation awareness and can result in excessive workload. In this article, we argue why the introduction of automated planners instead induces new human factors related problems such as complacency, opacity, and loss of situation awareness. In order to avoid such issues, we propose a mixed-initiative approach. The article describes our concept and the technical implementation of a mixed-initiative multi-vehicle mission planner. The planner serves as cognitive agent and supports a human operator during planning and re-planning processes. The article focuses on the interaction concept. A first experimental evaluation of the described interaction concept is presented. Our application comprises the teaming of manned and unmanned helicopters in complex military missions.

Keywords: AI systems · Associate systems · Human agent teaming · Mixed-initiative · Problem solving

1 Introduction

The mission planning and re-planning of multiple vehicles by a single operator under time constraints can result in excessive mental workload (MWL) conditions. Algorithms for automated planning and scheduling were developed that can solve such complex problems in reasonable time. However, these algorithms are often not directly suitable for use in incremental, user-centered collaborative planning [1]. Rather, the usage of such automated planners may result in loss of competences [2] and loss of plan awareness and plan comprehension. Additionally, most automated planners require programming exactly then when workload is already increased due to changes of the situation [3] which results in even more workload. Finally, the usage of fully automated planners may result in an inversion of hierarchy during mission execution as the human operator is obliged to execute a fully automation-generated plan.

In order to counteract such human factors issues, we propose a cooperative planning approach between human operators and cognitive agents based on incremental planning. The collaboration and cooperation of human operators with cognitive agents to solve a common planning problem is known as *mixed-initiative* (MI) planning. Although multiple mixed-initiative approaches were already presented in [4–6] the introduction of such systems in real-world applications is still missing. The challenge is

now to integrate such a system into complex multi-agent human-machine systems. Thereby, in order to bring an actual benefit, the focus of research must address an efficient interaction concept between human and agent. The design of the agent's intervention policy is of essential significance.

In this article, we present a mixed-initiative planning associate for helicopter onboard multi-aircraft mission (re-)planning. The agent is specifically designed to reduce automation-induced errors and to enable a helicopter pilot to solve multi-vehicle planning problems with sufficient quality in reasonable computation time. The proposed planning associate monitors the tactical situation and the pilot and intervenes during the planning process whenever necessary. The planning agent is integrated in a full military two seated helicopter mission simulation. The purpose of this article is to describe our concept, the implementation and evaluation of the planning associate. Thereby, the article focuses on the interaction aspects of the system.

2 Related Work

Allen and Ferguson present the design of a mixed-initiative agent that collaborates with a human operator in order to solve a common planning problem [6]. Thereby, they describe an integrated framework of several planning and reasoning functions and give a short evaluation. In [7] Chen et al. present experimental results on a mixed-initiative agent named RoboLeader which helps a human operator to coordinate a team of multiple UxVs. Roth et al. [8] describe the evaluation of a mixed-initiative system for multi-UAV mission management. The evaluation focuses on human-factors questions such as the measure of situation awareness, workload, and plan comprehension. Further research can be found in [9].

3 Application Manned-Unmanned Teaming

Our research application comprises the teaming of a manned two-seated helicopter with multiple unmanned aerial vehicles (MUM-T) in military helicopter transport missions. Characteristics of these missions are reduced preplanning time, highly dynamic mission environments (i.e. rapid changes of the tactical situation) as well as landings in hostile territory. Thereby, the unmanned systems are used as detached sensor platforms to increase the sensor range of the manned helicopter and to meet information demands for the pilots. The UAVs are used to reconnoiter the primary route of the manned helicopter and to find alternatives routes. Furthermore, the UAVs are able to locate suitable landing points in hostile territory. In critical mission phases, such as approach, ground operations and departure of the helicopter, the UAVs can provide protection. In this setup, the pilot non-flying is fully responsible for the tactical planning and re-planning of the manned/unmanned team during the mission. Mission planning tasks include helicopter route planning, contingency planning, and UAV task planning. The pilot uses a tactical map display to sketch plans and to command tasks to the unmanned systems. Thereby, he is assisted by our mixed-initiative planning agent, in order to increase performance and to reduce workload and human factors related issues. The agent proposes

reconnaissance tasks, such as the identification of targets, and task assignment to an aircraft depended on available resources. It identifies flaws and conflicts in the current plan and offers solutions. Furthermore, it helps optimizing a given plan. The interaction between pilot and agent is dialog-based using either text boxes on the tactical map display or voice interactions. However, the proposed concept is not restricted to our domain. Rather, it can be transferred to various applications with single or multi-vehicle planning by a single operator, such as the air traffic domain.

4 Concept

4.1 Work System

We define mixed-initiative as cooperation between human and agent, in which both can take initiative over the planning process and direct the process. Thereby, mixed-initiative systems can include adaptable and adaptive components. Other definitions can be found in [10–12].

The conceptual design is presented in Fig. 1 in work system notation [13]. The work system notation differentiates between the worker on the left-hand side and the tools on the right-hand side. The worker has knowledge of the overall work objective and tries to reach that objective by own initiative. He is furthermore authorized to change this work objective by own initiative. To achieve that work objective, the worker uses given tools which are shown on the right hand side of the work system. The tools are subordinates, i.e. hierarchically degraded with respect to the worker.

In our application, the worker is represented by a single human pilot. Multiple vehicles and their planning interfaces and algorithms, for example semi-automated

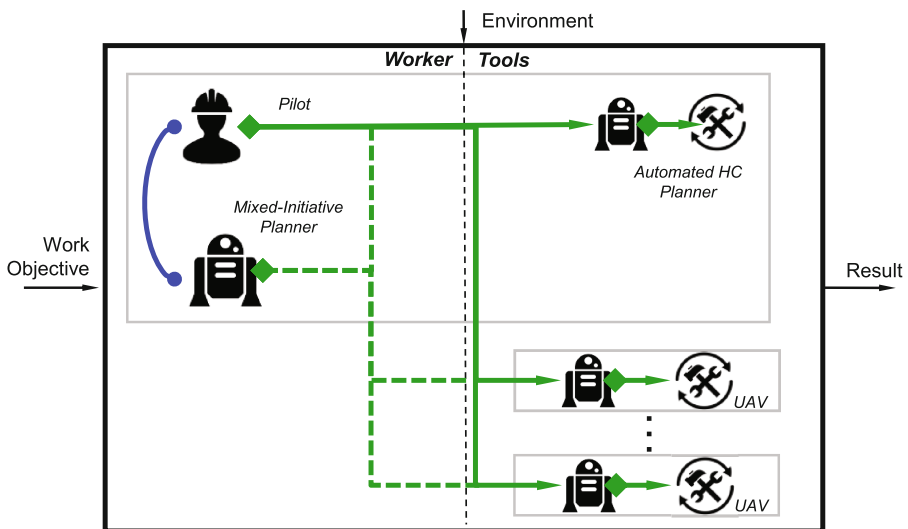


Fig. 1. Design pattern in the work system notation [13]

planner, serve as tools to the pilot. A cognitive agent onboard each unmanned aircraft (UAV) serves as delegate agent. Each delegate agent is controlled by the human worker using a task-based-guidance approach [14]. The agent controls the underlying conventional automation in a supervisory control relationship. The introduction of such a delegate agent reduces the workload of the human worker.

Finally, we introduce a cognitive agent as worker on the left hand side which re-presents the mixed-initiative planning associate. This agent has all characteristics of a human worker. That means, it is also aware of the mission objective and pursues it by own initiative. However, in contrast to the pilot, he is not allowed to define or modify the work objective. The agent has knowledge about the mission domain, available tools and given resources as well as the human pilot. The task of the cognitive planning agent is to assist the pilot in all situations which require mission planning and re-planning.

4.2 Behavior Rules of the Mixed-Initiative Agent

The task of the agent is to enable the pilot to solve multi-vehicle planning problems with sufficient quality in reasonable time. The agent shall increase planning performance and reduce workload of the human-agent team. Thereby, the agent shall help to mitigate human factors problems such as reduced situational awareness and reduced plan comprehension. In order to achieve these goals, we formulated the following behavior rules for the mixed-initiative agent:

- leave as much work to the pilot as possible,
- intervene as late as possible,
- intervene as little as possible, and
- intervene incremental, rather than complex, whenever possible.

In order to determine which information is required for the pilot at a given time we considered psychological aspects of human pilots. In organizational psychology in the area of planning and decision making Herbert Simon proposed the “Satisficing Principle” [15] which describes the behavior of decision makers. The principle underlines that the human in general does not try to find an optimal solution to a given problem. Instead, he stops working on the problem as soon as the solution is sufficient to his personal level of aspiration. We transferred Simon’s principle into our concept. Our agent does not try to reach an optimal solution. Rather, it stops intervening as early as possible. If the agent intervenes, it guides the pilot through the problem step by step. Thereby, it is designed to leave as much work as possible to the pilot.

Simon states that the human’s level of aspiration depends on the current situation. This means that in critical situations a low-quality solution is sufficient. In contrast, in very uncritical situations, the personal aspiration level of the human pilot might be much higher. We mapped these levels of aspiration into situation criticality and workload. If the pilot’s mental workload is excessive, his aspiration level might be much lower compared to a low workload situation. Similar, if the pilot finds himself in a critical tactical situation, his aspiration level is probably reduced. These assumptions result in further behavior rules for the agent:

- aspire after a sufficient plan, rather than an optimal plan,
- adapt the aspiration level to the current tactical situation, and
- adapt the aspiration level to the current human mental workload.

4.3 Interaction Concept

In the following, we describe the concept for interaction between human pilot and cognitive planning agent: in a first step, the pilot provides information about a mission goal to our cognitive planning agent. In the following course, the pilot uses planning tools to (re-)plan or to modify the mission plan whenever required. Figure 2 visualizes such a planning process; *initial state* and *final state* of the planning problem are highlighted. A plan is defined as a sequence of *actions* (*operators*) which transforms the *initial state* into the *final state*. In general, for a given problem multiple solutions exist. The *partial plan* denotes the currently implemented part of the full plan. Constraints, such as maximum route distance, minimum fuel on board, or time of arrival, can reduce the solution space. The pilot needs to find a solution, which does not violate any constraints.

The planning agent monitors the pilot and the situation and intervenes if required. There are three reasons for intervention:

1. The pilot does not continue with planning activities, even though it is required: If next planning steps are required and the pilot does not take actions, the agent informs about the problem and proposes next planning steps incrementally.
2. The pilot plan is erroneous: if (mission specific) constraints were violated in the implemented plan, the system informs about consequences and proposes alternative options.

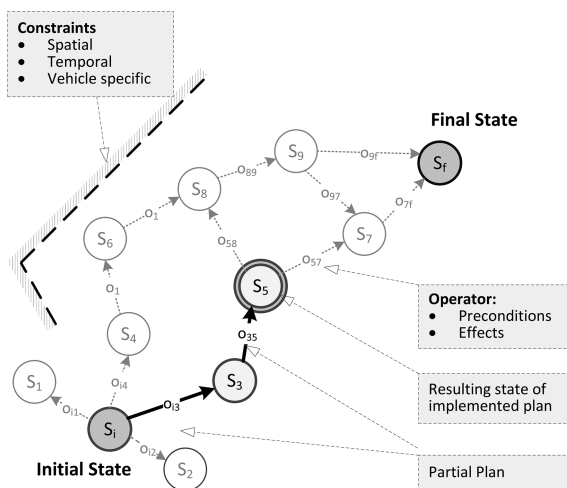


Fig. 2. Visualization of the planning process starting at an initial state

3. The pilot plans suboptimal: if the plan can be significantly improved, the agent proposes enhancements. However, since a valid solution already exists, the agent shall intervene only if the aspiration level of the pilot is high, i.e. the situation is uncritical and the pilot's workload is low.

The determination of the pilot's activities is required to shape the intervention policy. If the pilot is about to plan the helicopter route, we do not need to inform him about the necessity for planning. Furthermore, if the pilot works on the helicopter route problem, the agent should contribute to rather this problem than to less related problems.

Not only the agent has the possibility to initiate a dialog with the human, but also the pilot is able to initiate a dialog with the planning agent. The pilot can assign tasks to the automated planning associate in high workload situations or request information about future options and consequences. For example, he can ask the agent for optimizations of his mission plan. This allows human pilot to interact on own initiative with the associate agent which marks a fundamental difference to previous work.

4.4 Functional Architecture

In order to work with a team member on a mutual problem, first of all, both must identify the mutual goal. Secondly, a human team member must have knowledge about the work domain, activities of team members, and their mental states. In order to develop a cognitive agent, we transferred these key capabilities to our concept. We identified four system pillars which serve as key enablers for our mixed-initiative planner:

1. planning and plan reasoning,
2. pilots' activity determination,
3. pilots' mental plan progress assessment, and
4. intervention.

The first key enabler is the capability to plan and reason about plans. This capability is required to infer options and determine next planning steps for incremental planning. Furthermore, this pillar is required to reason about human generated plans, their shortcomings, and conflicts. Therefore, the knowledge about goals and planning domains is required. The second key enabler is the ability to determine the pilot's activities. Knowledge about the current pilots' activities is required to enable planner interventions which do not disturb the conversation flow. The third key enabler is a human mental plan progress assessment. Such an assessment is the accumulation of the pilot planning activities over time. If the pilot for example has noticed a plan conflict in the past, but not reacted to it so far, we do not need to inform him about the conflict later again, but maybe better support his reaction by any means. Finally, the fourth key enabler is the intervention generation component. This component forms the bidirectional interface to the pilot and interacts based on the behavior rules (Sect. 4.2). This component receives data from the three previously described components, generates an interaction strategy, and interacts as required to influence the environment. These four

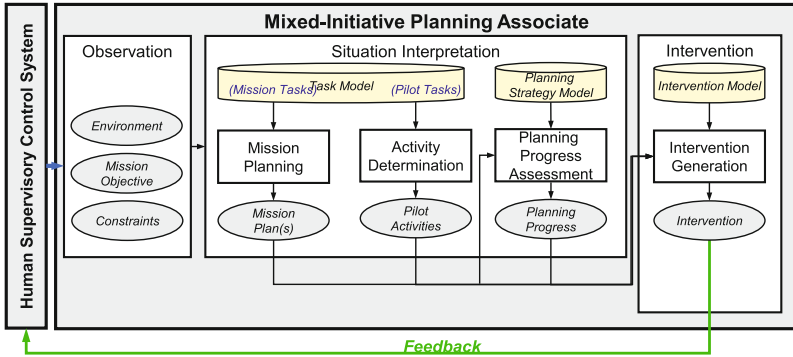


Fig. 3. Functional Architecture of the Mixed-Initiative Associate

pillars allow for an incremental and mixed-initiative planning process. Figure 3 shows the corresponding functional architecture of the agent. Our system can intervene on two different levels of automation. On the first level, it plans incremental and therefore requires more user interaction. On the second level, the agent intervenes with complex interactions rather than incremental. Thus, less user interaction is required. On the one hand, this reduces the mental workload of the pilot. On the other hand, it may also reduce pilot plan awareness. These two levels of automation can be adjusted by a workload-adaptive associate system, described in [16].

5 Implementation

This chapter describes the implemented system. The first subchapter shows the system architecture of our planning agent. The second subchapter describes the implemented HMI. The overall system architecture is presented in Fig. 4. The figure shows the important components of our agent on the left side as well as the implemented parts of the HMI and other relevant components on the tool side.

5.1 Agent System Architecture

Mission Planning. The first component, mission planning, represents planning and plan reasoning capabilities. It requires substantial domain knowledge in the area of military mission planning because otherwise the agent cannot assist the pilot adequately. For this reason, we conducted knowledge acquisition experiments with German military helicopter pilots to model our planning domain. The domain contains a model of tools, i.e. the manned helicopter and the unmanned systems. For logical planning and re-planning purposes, we modelled our planning problem in PDDL which is an action-centered language to solve planning problems [17]. Core of PDDL are actions with pre- and post-conditions that describe the applicability and the effects of actions. Figure 5 shows a graph-based visualization of the domain and their implemented

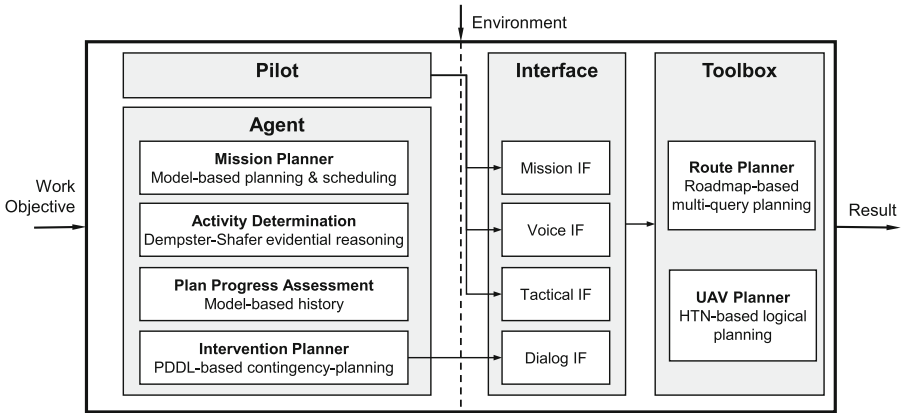


Fig. 4. Components of the implemented system in work system notation

actions. These actions comprise helicopter and UAV specific actions. We use a PDDL planner which works based on our mission domain and a problem file. Furthermore, a CPLEX planner is used for rapid task assignment, optimization and scheduling [18].

On runtime, information about capabilities of the UAVs (i.e. sensor equipment, transit speed, reconnaissance speed) is provided to the planner. Thereon, the planner generates

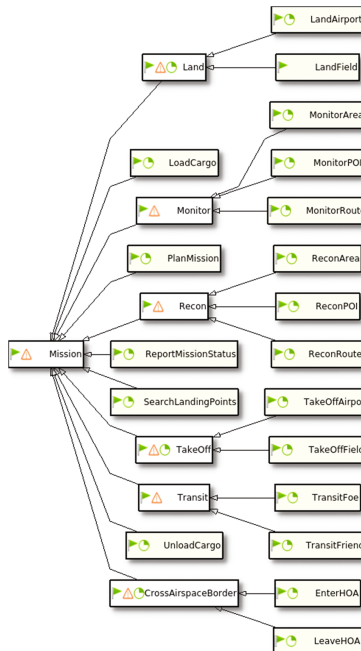


Fig. 5. Mission planning domain

a mission plan, based on the mission goal, the information demands by the pilots and available resources. However, usually the demand for reconnaissance information outreaches the available resources. Then, the planner prioritizes demands based on their criticality to the mission success. Most important are the detection of landing points, the protection during ground operations, approach and departure, as the helicopter is most vulnerable in these flight phases. Less critical are information demands, concerning alternative route segments. The generated mission plan is not directly forwarded to the pilot.

Activity Determination. The second component is the pilots' activity determination. Therefore, the pilot's interactions with automated cockpit functions are observed. Interactions can be manual or visual. In the second step, these observations generate evidences for hypotheses, which represent certain activities. Finally, these evidences are combined using the Dempster-Shafer theory. The result of the reasoning is the continuous determination of the pilot's current activities. The detailed concept and realization of the activity determination is presented in [19].

Plan Progress Assessment. The third component is called mental plan progress assessment and represents the pilot's mental state regarding the planning process. Therefore, it gathers data provided by the activity determination and accumulates the data to a mental plan progress. The gathered data comprise pilot helicopter planning activities and UAV planning activities.

Intervention Generation. The fourth pillar deals with the intervention generation. We modelled the interaction domain in PDDL as well. Interaction planning is handled using a logical contingency planner. The next sub-chapter will discuss the implementation of our intervention component in more detail. The purpose of the cooperative planning approach is to generate a mission plan which satisfies all constraints and is agreeable by the human pilot as well as by the agent. The purpose of the agent's intervention generation component is to derive an intervention strategy, i.e. a sequence of actions which can be used to transform an insufficient mission plan into a plan which is agreeable by all team members.

To enable cooperative behavior, we need to model the agent's behavior. We modelled an interaction domain in PDDL. The domain contains the environment and actions which can influence the environment if applicable. The environment consists of three pillars:

1. the specification of the tactical environment,
2. the description of the pilot's mental state, and
3. the description of mission related facts, such as mission plans, flaws, alternatives and optimizations.

Furthermore, the domain contains actions. An action can be applied to influence the environment if the preconditions for this action can be met. The actions represent all possibilities of the cognitive agent and can be used to determine the interaction strategy. Four different types of actions enable the planner to generate such a *mixed-initiative* interaction strategy. Types of actions are:

1. **Prioritizations** to determine the most important issue. For example, an enemy unit that threatens the mission plan must be prioritized higher than an optimization of the mission plan regarding the task assignment. The pilot’s activities influence the prioritization.
2. **Observations** to wait for further activities of the pilot. For example, the system can decide to observe the pilot’s next interaction with the tactical display before intervening.
3. **Pilot interactions** to communicate with the pilot. This type of interaction is most important for cooperation between agent and pilot. The interaction of the agent is dialog-based. It includes informative dialogs, proposals for future routes and target candidates for reconnaissance or proposals for optimizations.
4. **System interactions** to modify the state of a technical system. These include the modification of tools such as the route planner or the UAV-planner (Fig. 4).

On each simulation step, the interaction planner determines the most important planning problem and then generates a proper handling strategy to solve the problem. As described in the previous chapter, it is most important to avoid flooding the pilot with information. Rather, only mission relevant information and proposals shall be communicated to the pilot. This represents Simon’s Satisficing principle which is modeled as an optimization function. An example of such a generated interaction strategy is displayed in Fig. 6.

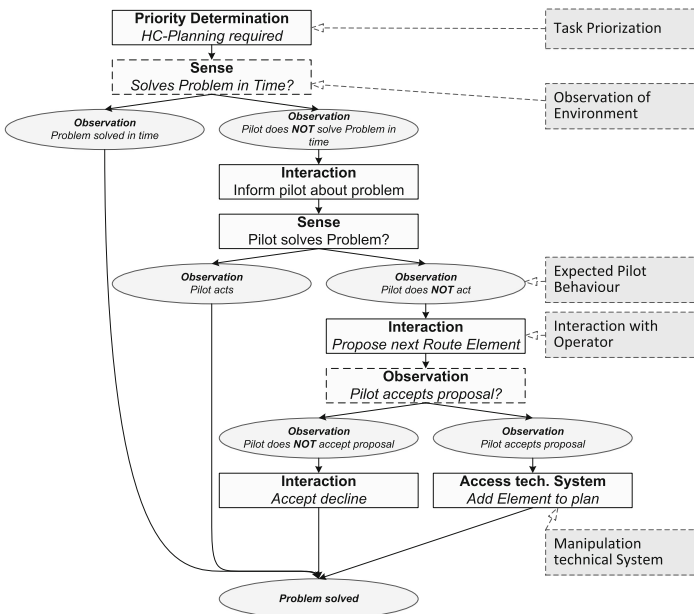


Fig. 6. Interaction strategy generated by the contingency-based intervention planner during runtime

5.2 Pilot Planner Interface

We developed an HMI that satisfies requirements for mission planning and communication between pilot and agent. The HMI has the following components (also displayed in Fig. 5):

1. a mission interface which is used by the pilot to specify the mission objective,
2. a tactical map display, which is used by the pilot to sketch the mission plan and command tasks to the UAVs (this display can be also used by the agent to visualize plan proposals and alternatives),
3. a dialog interface for the agent to communicate with the pilot using text boxes, and
4. a dialog interface for the pilot to initiate a communication directly with the agent.

The tactical map display is shown in Fig. 7. The left side shows the pilot's interface (an object-oriented context menu) to sketch a mission plan and command tasks to the UAVs. The right side shows an agent initiated dialog to solve a problem. The text box on the upper right is used by the agent to explain the problem and to point out a solution. Additionally, the solution is visualized on the map display in magenta.

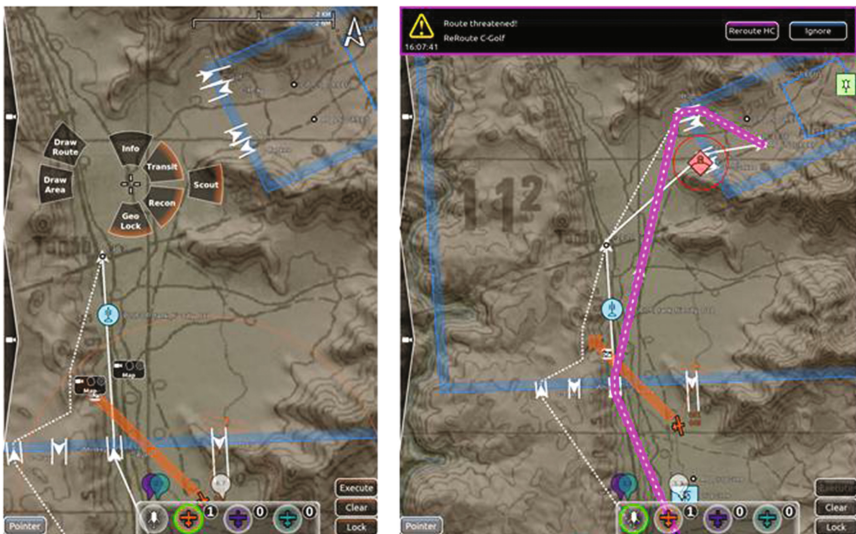


Fig. 7. IFS Helicopter simulator cockpit

Furthermore, we developed a pilot-agent dialog voice interface. This interface allows the pilot to interact on own initiative with the associate agent which marks a fundamental difference to our previous work. The interface can be used to request information and proposals from the agent and to add constraints to the mission plan. Therefore, we implemented the following grammar for commands:

- Commands: [Keyword] [Affected Vehicle] [Command] [ContextVariable]
- Requests: [Keyword] [Request] [Affected Vehicle] [ContextVariable]

The implemented interface is based on the software SIMON which is an open source front-end for speech recognition.

6 Experimental Evaluation of Agent-Human Interactions

In an experimental campaign, we examined the impact of our mixed-initiative agent-human interactions on the pilot's workload and the overall planning performance. We hypothesize that during a planning process:

- plan segment proposals of the mixed-initiative agent reduce the workload of the pilot, and
- planning performance can be increased using plan segment proposals.

6.1 Configurations and Research Setup

In order to examine the hypotheses, a comparative experiment was conducted with four different configurations.

- **Configuration A:** The planning problem is simple (1 UAV, 3 Targets). The pilot uses a tactical map display to task the UAV manually.
- **Configuration B:** The planning problem is simple (1 UAV, 3 Targets). The planning agent proposes a task assignment. The pilot has to accept or decline the proposal after verification.

Our experimental hypotheses state that configuration B increases performance and reduces MWL compared to configuration A. The performance was operationalized using the time required to fulfill a given task. More specifically, based on a task with a given complexity, the performance is operationalized by the inverse of the time used to execute the planning task. The MWL construct was operationalized using the NASA-TLX [20] questionnaire as a subjective measure.

As experimental design we chose a within-subjects design. Each subject had to perform each configuration 5 times. The sequence of configurations was randomized between all participants in order to eliminate sequence effects.

In order to enforce adequate proposal verification by the subjects, one of the agent's proposals was incorrect for each subject. The subjects were informed about the possibility of incorrect proposals.

6.2 Participants and Experimental Conditions

The experimental test sample consisted of 21 cadets of the German Armed Forces. The participants aged between 18 and 29 years ($M_{\text{age}} = 22.6$, $SD_{\text{age}} = 2.2$). During the experiment the participants planned one UAV on a single tactical map display. The display showed the UAV and its equipment as well as possible targets. All subjects were trained with the software previously to the experiment.

6.3 Results

Experimental data set was generated in [21]. Figure 8 shows the results for the workload and performance evaluation. The assessment of overall workload showed differences between configurations A and B. Compared to configuration A ($M = 32.12$; $SD = 16.91$), the workload could be reduced by 28.18% in configuration B ($M = 23.07$; $SD = 11.20$, $p = 0.053$). The assessment of overall performance showed significant differences between configurations A and B. The time required to perform the simple planning task manually ($M = 15.3$; $SD = 5.97$) could be reduced with assistance in Configuration B ($M = 10.08$; $SD = 4.63$; $p = 0.032$) by 34.14%. The results show that both hypotheses can be accepted.

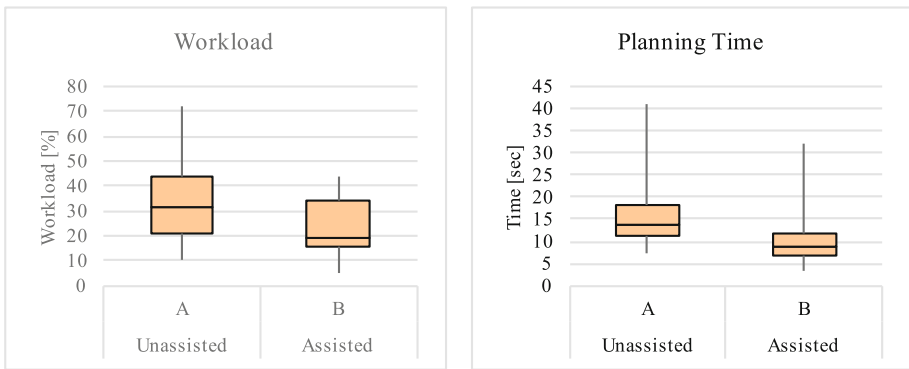


Fig. 8. Results for workload and planning time for configurations A and B

7 Experimental Evaluation of Human-Agent Interactions

In order to evaluate the effectiveness of the developed human-agent interaction concept, we conducted another experiment in our helicopter mission simulator. The experiment was conducted to evaluate if pilot initiated dialogs with the mixed-initiative agent can reduce the workload and increase the performance of the human-agent team.

Therefore, we developed an experiment in which a single pilot has to fly a simplified MUM-T mission and is additionally responsible for helicopter route planning and UAV task assignment. We hypothesize that the pilot flying the helicopter benefits from the speech interface. Furthermore, we hypothesize that pilot initiated interactions with the mixed-initiative agent will affect the work result of the system in a positive way and increases performance and reduce mental workload of the pilot.

7.1 Configurations and Research Setup

For the evaluation of the system, a comparative experiment was conducted, in which three missions (*mission I, II and III*) were performed. In the experiment we evaluated

the described interface of the mission planner. Thereby, we compared three configurations:

- **Configuration A:** the planning process is executed using a tactical map display without speech interaction. The pilot works with his tools using manual interactions.
- **Configuration B:** Planning is done using basic speech commands, which are equal to the planning commands in Configuration A. The pilot works with his tools using speech interactions.
- **Configuration C:** Planning is done with advanced speech interaction directly with the mixed-initiative agent. The pilot works with the planning agent using speech interaction.

The experimental hypothesis says that configuration B increases performance compared to configuration A; configuration C increases performance compared to configuration B. Furthermore, it can be assumed that Configuration B reduces workload compared to Configuration A and that configuration C reduces workload even more. In the experiment, all participants were exposed to the three configurations of the system. Therefore, the experimental design is a within-subjects design. In order to derive mental workload of the pilot, we added a secondary task. To eliminate sequence effects, we conducted the experiment with two test groups and switched the sequence of the configurations between both groups. Test group one conducted the configurations in ascending order. For test group two we reversed the sequence.

The missions were designed as follows. The subject had to fly a helicopter at an assigned altitude and speed in our helicopter mission simulator. The simulator cockpit is shown in Fig. 9. During each flight five re-planning situations occurred. To re-plan the mission, the subjects could use either the tactical map display or voice interaction, as according to the experimental condition under evaluation. To ensure comparability, all missions had an identical layout.

In order to prove the hypotheses, the constructs (MWL and performance) were operationalized using the following dependent measures. We operationalized MWL



Fig. 9. IFS helicopter simulator cockpit

using a secondary task. Here, the subject had to classify possible targets on a display, whenever possible. To determine the workload of the primary task, we evaluated how much interaction time was spent in the secondary task similar to [22]. We operationalized performance as deviation in altitude and speed from the intended flight path.

7.2 Participants

The sample consists of 10 persons recruited from the University of the Bundeswehr Munich. Participants include 8 officers of the German Armed Forces and 2 Engineers. Participants include 9 male and one female. The participants aged between 22 and 31 ($M_{\text{age}} = 27.2$, $SD_{\text{age}} = 2.6$).

7.3 Results

Results for the secondary task are presented in the following. The average time required to classify an object in configuration A ($M = 14.7$; $SD = 13.47$) could be reduced in configuration B ($M = 9.4$; $SD = 5.66$) and configuration C ($M = 7.8$; $SD = 5.4$). Figure 10 shows the results for the performance increase in the secondary task. The graph shows the averaged performance increase. Thereby, configuration A serves as a baseline measure. The averaged performance in configuration B is 25% higher and in configuration C 40% increased, compared to configuration A. These results indicate that the workload in the primary tasks could be reduced. Therefore, the hypothesis that states that MWL can be reduced using voice interactions can be accepted.

Figure 11 visualizes results of the increase of performance in the secondary task in boxplot format for configurations B and C with regard to configuration A.

The evaluation shows that the standard deviation for altitude, referred to the overall mission, could also be reduced, displayed in Fig. 12. The standard deviation in altitude could be reduced in configuration B by 40.7% and in configuration by 41.7% compared to configuration A which served as reference. This shows that voice interaction in general could increase planning performance. However, the direct agent interactions could not increase performance further.

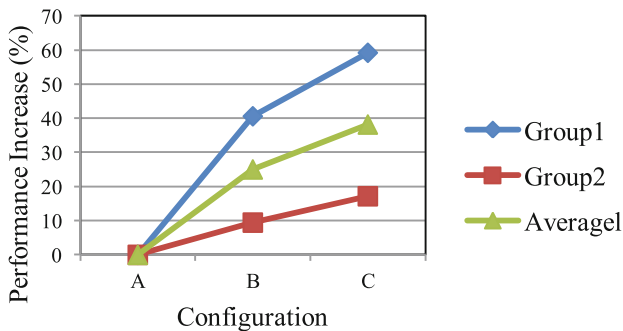


Fig. 10. Results of performance increase for the secondary tasks for all three configurations

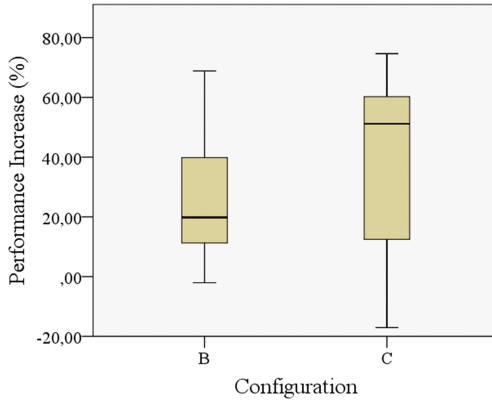


Fig. 11. Results of performance increase for the secondary task for configurations B and C

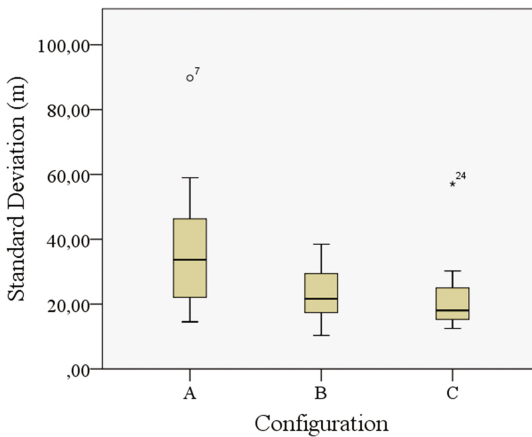


Fig. 12. Deviation from specified altitude over configurations A, B, C

In general, it must be said that the implemented speech recognition rate was depended on the subject. The results show that workload could be reduced using voice interaction instead of manual display interactions if the pilot is flying the helicopter. However, our experience shows that if the pilot is not flying the aircraft, manual display interactions are preferred.

8 Conclusion

This article describes our concept for a mixed-initiative planning agent, integrated in a two seated military helicopter mission simulation. The agent shall ensure correct mission planning and re-planning of multiple unmanned vehicles in high workload conditions. The pilot and the cognitive agent cooperate and solve mission (re-)

planning problems incrementally. Thereby, the cooperation between pilot and agent is dialog-based (text or voice). During runtime, our agent monitors the pilot's activities, as well as the tactical situation and the given partial plan. Based on the currently implemented partial plan, the current tactical situation and the pilot's activities the agent generates an interaction strategy which is used to assist the pilot. The evaluation shows that the implemented interfaces reduce workload and increase performance in onboard re-planning situations. Further research is required in order to demonstrate the benefits of the agent's interaction strategy. We are currently in the preparation phase for full mission experiments with German military helicopter pilots.

References

1. Allen, J., Ferguson, G.: Human-machine collaborative planning. NASA Plan. Sched. Work (2002)
2. Wiener, E.L., Curry, R.E.: Flight-deck automation: promises and problems. *Ergonomics* **23** (10), 995–1011 (1980)
3. Wiener, E.L.: Human factors of advanced technology ('Glass Cockpit') transport aircraft. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, June 1989
4. de Brun, M., Moffitt, V.: Mixed-initiative adjustable autonomy for human/unmanned system teaming. In: AUVSI Unmanned Systems North America Conference (2008)
5. Chen, J.Y.C., Barnes, M.J.: Human-agent teaming for multirobot control: a review of human factors issues. *IEEE Trans. Hum.-Mach. Syst.* **44**(1), 13–29 (2014)
6. Ferguson, G., Allen, J.: TRIPS: an integrated intelligent problem-solving assistant. In: AAAI/IAAI, pp. 567–572 (1998)
7. Chen, J.Y.C., Quinn, S., Wright, J., Barnes, M., Barber, D., Adams, D.: Human-agent teaming for robot management in multitasking environments. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 103–104 (2013)
8. Roth, E.M., Hanson, M.L., Hopkins, C., Mancuso, V., Zacharias, G.L.: Human in the loop evaluation of a mixed-initiative system for planning and control of multiple UAV teams. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **48**, 280–284 (2004)
9. Clare, A.S., Macbeth, J.C., Cummings, M.L., Member, S.: Mixed-initiative strategies for real-time scheduling of multiple unmanned vehicles. In: American Control Conference, pp. 676–682 (2012)
10. Nickerson, R.S.: On conversational interaction with computers. In: ACM/SIGGRAPH Workshop on User-oriented Design of Interactive Graphics Systems, UODICS 1976, pp. 101–113 (1976)
11. Walker, M., Whittaker, S.: Mixed initiative in dialogue. In: Proceedings of the 28th Annual Meeting on Association for Computational Linguistics, pp. 70–78 (1990)
12. Burstein, M.H., McDermott, D.V.: Issues in the development of human-computer mixed-initiative planning. *Adv. Psychol.* **113**, 285–303 (1996)
13. Schulte, A., Donath, D., Lange, Douglas S.: Design patterns for human-cognitive agent teaming. In: Harris, D. (ed.) EPCE 2016. LNCS (LNAI), vol. 9736, pp. 231–243. Springer, Cham (2016). doi:[10.1007/978-3-319-40030-3_24](https://doi.org/10.1007/978-3-319-40030-3_24)
14. Uhrmann, J., Schulte, A.: Concept, design and evaluation of cognitive task-based UAV guidance. *Int. J. Adv. Intell. Syst.* **5**, 145–158 (2012)
15. Simon, H.A.: Rational choice and the structure of the environment. *Psychol. Rev.* **63**(2), 129–138 (1956)

16. Brand, Y., Schulte, A.: Adaptive Assistenz für Militärische MUM-T Hubschraubermissionen. In: 5. Interdisziplinärer Workshop Kognitive Systeme: Mensch, Teams, Systeme und Automaten Verstehen, Beschreiben und Gestalten kognitiver technischer Systeme (2016)
17. McDermott, D., Ghallab, M., Howe, A., Knoblock, C.: PDDL-the planning domain definition language. In: AIPS 1998 Planning (1998)
18. Schmitt, F., Schulte, A.: Mixed-initiative mission planning using planning strategy models in military manned-unmanned teaming missions (2015)
19. Schulte, A., Donath, D., Honecker, F.: Human-system interaction analysis for military multi-rpa pilot activity and mental workload determination. In: IEEE SMC Conference (2015)
20. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Adv. Psychol.* **52**(C), 139–183 (1988)
21. Rump, H.: Experimentelle Analyse von Leistungsfähigkeit und Beanspruchung eines Operateurs bei der Multi-UAV Aufgabenzuweisung [Experimental Evaluation of Operator Workload and Performance during Multi-UAV Task Assignment] (Bachelor Thesis). Universität der Bundeswehr, Munich, Germany (2017)
22. Crandall, J.W., Goodrich, M.A., Olsen, D.R., Nielsen, C.W.: Validating human-robot interaction schemes in multitasking environments. *IEEE Trans. Syst. Man, Cybern. Part A Syst. Hum.* **35**(4), 438–449 (2005)

A Field Study of Multimodal Alerts for an Autonomous Threat Detection System

Erin T. Solovey¹, Pallavi Powale², and M.L. Cummings³(✉)

¹ Drexel University, Philadelphia, PA, USA
erin.solovey@drexel.edu

² Google, Mountain View, CA, USA
pallavipowale00@gmail.com

³ Duke University, Durham, NC, USA
mary.cummings@duke.edu

Abstract. Every year, inattentive or impaired drivers strike law enforcement officials, emergency personnel, and other workers by the roadside. Preventative efforts include making at-risk parties more conspicuous to oncoming motorists in order to prompt safer driving behaviors. In contrast, this work evaluates active alerting mechanisms designed to induce defensive action from at-risk roadside personnel once a hazardous situation has been autonomously detected. This paper reports on field investigations with state police to capture their cognitive requirements for this dynamic environment, as well as the design of four alert prototypes for a high noise, low-light environment such as a highway shoulder. We discuss implications for such future autonomous systems and argue that such active defensive alert mechanisms could improve roadside safety and save lives.

Keywords: Autonomous alerting, safety · Law enforcement · Wearable computing · Ubiquitous computing · Haptics · Multimodal alerts

1 Introduction

In the United States between 2000 and 2009, 120 law enforcement officers were struck and killed by vehicles while performing duties such as directing traffic, assisting motorists, or stopping on a highway shoulder [24]. For example, on October 19th, 2012, a Nassau County highway patrol officer exited his vehicle on an expressway to investigate a crash and aid an injured person. Soon after, he was struck and killed by another car [4]. On December 29th, 2012, an interstate officer was hit and knocked over a guardrail after stopping another vehicle on the highway [23]. In June of 2013, despite emergency lighting, a Massachusetts state trooper was rear-ended by a drunk driver while making a roadside stop [8]. Had he been outside the car, the collision could have been fatal. Similar traffic threats affect other first responders and roadside personnel as well. In 2008, 29 of 114 firefighters killed on duty in the U.S. were killed in vehicle accidents. Between 1992 and 1997, at least 67 EMS providers perished in ground transportation-related events [25].

While numerous precautionary steps can be taken to protect roadside personnel, drivers are fallible [12]. Unlike most other passive mitigations developed to warn

drivers of the presence of roadside personnel, this paper explores an active defensive alerting system for at-risk individuals, which incorporates technology to communicate danger.

Automatic identification of anomalous driving behavior is becoming feasible as sensors and computer vision systems improve. An on-cruiser, backward-facing computer vision system has been proposed to monitor oncoming traffic for police officers stopped roadside [13]. The system works by means of two cameras processing video at high and low resolution to track vehicle trajectories as they approach the cruiser. If the system detects a dangerous vehicle trajectory, an imminent danger signal could be delivered to the officers (Fig. 1). This monitoring system is augmented by computer-controlled laser road flares to divert traffic [27]. Autonomous hazard recognition systems such as this could act as a line of defense when other passive signals fail. However, the design of the warning signal requires careful consideration of the unique human factors for this user group.



Fig. 1. Chain of communication from autonomous detection of dangerous vehicle to user notification

This paper describes the design and testing of alerts that can communicate the danger recognized by a machine vision system to a person at risk. It was designed to be easy to use, unambiguous, and efficient in mobilizing the user to take preventive action. It was also important that the design was technically and fiscally feasible, and that potential users were willing and inclined to use this mechanism. This work has the potential to save hundreds of lives and provide a more effective alternative to existing alerting mechanisms.

The contributions of this paper are as follows. First, we report on the results of a field investigation to characterize police officers' operational environment, focused on highway safety. We then introduce four personal defensive alerting devices for such an environment as well as an evaluation experiment to assess their relative effectiveness. Finally, we provide insights on the practical implementation of defensive alert systems from a focus group, which included state police officers.

2 Background

2.1 Hazardous Highway Conditions

The factors that contribute to hazardous traffic conditions are manifold. First, motorists can be distracted while driving or be unfit to drive. In 2010, there were a total of 195,879 Driving Under the Influence (DUI) arrests in California alone [6]. Poor highway engineering can endanger police officers and other personnel on the road as

well. The Arizona Crown Victoria Police Interceptor Blue Ribbon Panel and the New York State Police recommend that officers position their highway stops parallel to the highway and a sufficient distance from both violator vehicles and the edge of the highway [1]. Unfortunately, these types of stop locations are not always available. Highway engineers are often forced to reduce shoulder width or remove emergency breakdown lanes to help mediate high traffic volume [1]. In addition, highway traffic can be loud, visually demanding, and constantly changing. Weather conditions and terrain can reduce visibility of the surrounding area, making it harder to find escape routes, and temperature can cause discomfort and impaired tactile discrimination, especially in the cold [18]. These conditions impair an officer's ability to respond to threats, with little margin to escape an imminent collision.

2.2 Highway Safety Policy and Practice

A number of national agencies, including the American Association of State Highway and Transportation Officials and the National Safety Commission as well as international groups such as the International Association of Chiefs of Police, are working to minimize such hazards. In 2007, forty-three states had passed "Move Over" laws, which require oncoming traffic to clear the lane closest to a stopped officer [15]. However, the laws were not reliably enforced and in a survey taken that year, 71 percent of Americans reported no knowledge of these laws [15]. The Michigan Give 'em a Brake Safety Coalition supports the establishment of modified speed limits in work zones, and they have also campaigned through bumper stickers and over the radio [14].

2.3 Highway Safety Devices and Technology

In addition to policies, many types of visual, auditory, and haptic devices and technologies are used to raise driver awareness of unusual road conditions.

Visual signals. Traffic cones, flares, signs, message boards, and reflective markings are all used to control and divert traffic. Police uniforms often include retroreflective garments such as jackets and raincoats to help improve their conspicuity, and the Federal Highway Administration requires that such garments comply with American National Standards for High Visibility Safety Apparel and Headwear to ensure their visibility [11]. Additionally, studies have shown that retroreflective striping on police cruisers is particularly effective, including fluorescent colors during the day and contrasting colors to make objects stand out from the background [25]. Using LEDs, colors and light patterns can be varied based on the amount of ambient light [1].

Auditory signals. Sirens and horns, today a quintessential feature of emergency vehicles, exemplify the auditory modality of warning signals on the road. While valuable when cutting through traffic and effective at grabbing attention, these loud, conspicuous warnings can cause physical discomfort at close range, be obstructive to covert police work and unnecessarily disturb communities. For this reason, sirens are typically used only for brief periods, and rarely on stationary vehicles.

Haptic signals. There are also haptic methods currently in place for protecting against vehicle accidents. For example, Sonic Nap Alert Patterns (SNAPs) are indentations in the road surface that produce a loud noise and vibrations in a vehicle driving over it [28]. The use of SNAPs on the Pennsylvania Turnpike over five different projects produced a seventy-percent reduction in “drift off road” accidents. These types of haptic patterns, now more loosely referred to as “rumble strips” have also adapted to be raised features in plastic, ceramic, or asphalt materials, and have been used in various locations such as parking lots and between highway lanes. Rumble strips have also proved to be “more cost effective than many other safety features including guardrails, culvert-end treatments, and slope flattening [9].” These haptic mechanisms, however, are geared toward motorists and are permanent features installed on the ground. There has been little work looking at more dynamic mechanisms.

3 Field Investigation and Design Considerations

To address the safety issues of individuals conducting roadside stops, we conducted a field investigation in which we spent two evenings on “ride alongs” with state troopers during actual highway patrol [16]. This allowed us to identify important considerations for designing a defensive alerting system. In addition, our design was informed by previous research on alerts. Further details on the ride alongs and related research are described below.

3.1 Ride Alongs

Police officers are highly trained individuals, skilled in fast decision-making, safety procedures, and emergency response. They are trained to be familiar with their equipment and to be prepared for a wide range of situational circumstances. To gain a better understanding of their operational environment, we conducted two ride alongs with a sergeant from the State Police. We were interested in understanding the cognitive requirements of the job by observing officer behavior and in characterizing all other haptic, visual, or auditory stimuli the officers experienced, with the purpose of gauging the sensory load of the environment. Decibel readings were taken around the vehicle on the shoulder of the road and notes of equipment and uniforms were collected.

During the ride alongs, we sat in the passenger seats of a police cruiser and observed the officer at work. The ride alongs were conducted in late fall and after sunset, so the environment was cold and dark. We were given reflective jackets to wear as an additional safety measure. Over the course of each ride along, the officer stopped in four different roadside locations, both on the highway and in more suburban settings. At these stops, while the officer attended to the infraction and stopped party, we would exit the vehicle to collect data using digital cameras, video cameras, a decibel meter, and pen and paper. We also interviewed the officer before, during and after the ride along for clarification and further details.

Timeline of a Roadside Stop. From observations during the ride-alongs and an informal cognitive task analysis with the officer, we were able to gain an understanding

1.!Target!Dangerous!Motorist!
2.!Run!License!Plate!
3.!Select!Safe!Stop!Location!
4.!Activate!Siren!
5.!Make!Stop!and!Assess!Danger!!
6.!Approach!Vehicle!on!Foot!
7.!Take!Action!

Fig. 2. Sequence of events in a roadside stop. Officers are trained to take these steps while driving, attending to oncoming traffic, and planning for emergency situations.

of the timeline of a roadside stop. Figure 2 summarizes the sequence of events observed each time the officer conducted a stop.

A roadside stop occurs when an officer targets a motorist on the road. As an officer begins to tail a subject vehicle, the targeted motorist will usually know that he or she is being followed by a police officer. However, the identity of the motorist is unknown to the officer. Inside the cruiser, each police officer has a computer interface on which he or she can run the license plate to match plate numbers with the registered owner of the vehicle, potentially the dangerous driver at hand. It is possible, however, that the current driver is not the owner. The car may be borrowed, leased, or even recently stolen and not yet reported. Once the license plate has been analyzed on the computer, the police officer will select a safe location to make the stop. At this point, the police siren may be activated, and the officer will make the stop and assess danger. Sometimes the target vehicle's driver will comply and sometimes the driver will not. For example, the driver may panic and stop his vehicle in the middle of the road or on the opposite side of the highway where no breakdown lanes exist. The motorist may become hostile or try to escape the situation. This latter possibility becomes more of a concern the longer the vehicle takes to come to a stop. As he is driving, the driver may be readying a concealed weapon or searching for a personally advantageous location to stop where the officer's attention may be diverted.

The officer must then approach the vehicle on foot, from up to 100 yards away. At this time, his or her attention is always divided between the stopped individual and oncoming traffic, both of which can pose serious threats to safety. Officers are trained to always look for escape routes in their environment, make their presence known to oncoming traffic, but also to conceal themselves from the targeted driver. If and when the officer must move to a safe location, time is critical. For example, a car traveling at 70 mph will travel 100 yards in 2.9 s. Based on an assessment of the situation, the officer chooses between wearing reflective gear or standard jackets, plans movement around the stopped vehicles and will then write a citation or take other action. Most of this behavior is taught through training and practiced until it is routine.

Table 1. Summary of ride along sound levels

Source	Max reading (dB)
Inside vehicle	69.0
Highway shoulder	83.3
Suburban neighborhood	71.0
Horn	85.5
Air horn	90.5
Sirens (Wail, Yelp, Piercer)	92.9, 90.5, 90.7

Noise in Roadside Environment. Adecibel meter was used to measure sound levels in various locations over the course of the ride along. Readings were also taken of other warning signals currently in use. The results are summarized in Table 1. Outside the vehicle in traffic, maximum decibel readings varied between 71 and 84 dB, mostly from vehicles rushing past. Inside the vehicle, the readings reached up to 69 dB. The cruiser’s built-in sirens, gauged from about 30 feet away from the vehicle, reached decibel readings in the 90 s. All sirens are automatically turned off when the car is in park. The officer also carries an on-person radio and multiple other radios inside his vehicle.

Officer Uniform and Equipment. Uniforms consist of combat boots, a long-sleeved shirt and slacks (or shorts in the summer). All on-person items are carried on an external waist belt or cross-chest belt. Officers might wear several other layers of clothing (e.g. a vest, jacket, or undershirt) and their standard-issue equipment can include a variety of other items such as radios, cell phones, and firearms, for various situations.

3.2 Additional Design Considerations

Alert Perception. Humans are generally capable of selectively attending to individual channels of stimuli even in the presence of other competing stimuli [19]. However, an emergency alert that can tap into unengaged cognitive resources will have the best chance of capturing attention. For example, in a loud setting, one might choose to use a non-auditory alert for better chance of detection. But even an easily detected alert can be inaccurately identified, especially in a high-pressure environment or when used with other similar alerts. In light of these challenges, we can manipulate the content of the signal to optimize the user’s perception and response to the alert. For example, certain sounds may have preexisting connotations for humans that would accelerate their reaction time. A siren would more quickly and intuitively be identified as an oncoming emergency vehicle than a foghorn or a doorbell. Furthermore, physical characteristics such as frequency and volume can also enhance detection and reaction time. These considerations are further discussed below.

Alert Modalities. Roadside alerting requires a modality of warning that would be effective in all types of lighting and weather conditions and which would capture

attention as quickly as possible. We explored the characteristics of three different modalities of warning: visual, auditory, and haptic.

Visual alerts. On the road at night, traffic headlights can cause much light pollution and glare and at different times of the year, fog, frost, dew and dirt can also significantly degrade visibility [26]. Moreover, crash warning system guidelines published by the National Highway Traffic Safety Administration specifically recommend visual warnings for “continuous lower-priority information” and discourage their use for “conveying time-critical information” [20]. Considering the operation environment of the highway, we concluded that a visual alert would not be appropriate.

Auditory alerts. Auditory signals are effective as warnings because they act on a sense that is not easily ignored, and could “be detected automatically and routed through on a priority line to the brain [17].” Three different types of auditory sounds can be used as warning signals: abstract tones, auditory icons, and verbal messages [2, 7, 26]. An abstract tone is typically composed of a single or multiple tones, which can be pure or harmonically complex. These tones can be continuous, they can be pulsing, or they can otherwise vary temporally, but the distinct pattern of sound, whatever it may be, must be identifiable to humans and will require learning. It has been found that warnings that consist of single continuous tones or similar temporal patterns are easily confused [7].

Auditory icons are sounds that typically have pre-existing associations with the warning audience [10, 26]. They are typically composed of real-world sounds that have a relationship with the circumstances they represent. For example, using the sound of skidding tires to notify the user of a vehicle crash [26]. Because of this relationship, auditory icons are easier to learn and identify than abstract tones.

Finally, verbal auditory messages, like verbal visual messages, use language to signal warnings. They have similar benefits, costs, and challenges. Incoherent or long messages will delay reaction times and in a high noise environment. Verbal messages can also cause a language barrier when the user population speaks different languages. However, if the appropriate language is used, verbal messages require the least learning, which could be suitable for an infrequent warning or one that appears in stressful situations that might cause listeners to forget the meaning of a more abstract alert [26].

Aside from the content of the sound, the physical characteristics of the signal can also be used to manipulate perception. Research has shown that “fundamental frequency, harmonic series, amplitude envelope shape, delayed harmonics, and temporal and melodic parameters such as speed, rhythm, pitch range, and melodic structure all have clear and consistent effects on perceived urgency [3].” These characteristics also play an important part in the conspicuity and discriminability of the signal, two features that the National Highway Traffic Safety Administration has indicated are most important in the design of imminent collision warnings [20].

The human auditory system is much better at perceiving changes in sounds rather than absolute frequency or intensity [17] and furthermore, warnings sounds will generally be more resilient against environmental noise if they are composed of multiple sinusoidal tones [26]. Regarding alert amplitude, guidelines suggest that high urgency

warnings should be 10–30 decibels higher than the masked threshold, a measurement of listener hearing threshold based on frequency and decibel level [20, 26].

Haptic alerts. Haptic warnings have not been studied or used as widely as auditory alerts in roadside environments. However, touch is an underutilized sensory channel and research has shown promising prospects for haptic alerts in comparison to visual and auditory warnings. In a study on collision avoidance, it was observed that reaction times to rear-end collision warnings was significantly shorter using tactile warnings than using visual warnings in a simulated driving environment and potentially also shorter than auditory warnings in real driving situations [21]. Rumble strips have now been installed all over the United States have drastically reduced drift-off-road accidents [9, 26, 28]. Although the roadside environment for officers and other workers is different than that for drivers in their cars, these studies suggest that tactile cues can be useful in circumstances that are perceptually taxing on the visual and auditory system.

Auditory or visual cues are often better indicators of orientation and location. They are both distal senses, capable of containing information about the distance of an event [22]. However, if auditory and visual cues are impractical, haptic alerts can also be used to orient attention using directional spatial tactile cues. In a study, drivers were warned of front-end collisions through a haptic vibration on the stomach and rear-end collisions through a vibration on the back. In cases where directionally appropriate cues were given, responses were 66 ms faster [22].

Haptic warnings are recommended in conjunction with warnings of other modalities to present redundant information [5, 20]. The combined message can create a sense of enhanced importance and enlarge the audience for which the warning will be effective, for example, persons with disabilities in perceiving other modalities [26].

4 Alerting Mechanism Requirements

Based on knowledge gathered from the ride along and the literature review, the requirements for the warning system were finalized, and are listed below.

- (1) For the best chance of detection, the alert must excite a sense that is not otherwise engaged or over stimulated in the operational environment.
- (2) The alert must produce the desired effect in less than 3 s. It is crucial that the speed of hazard detection and communication to the officer is maximized.
- (3) The alert signal must be succinct but descriptive enough to trigger both fast and accurate recognition.
- (4) The alert must be more urgent than, and distinct from, other signals the officer may already have in use.
- (5) The alert must be effective up to 100 yards away from the police cruiser, since this is a typical distance officers travel from their car.

In addition to these technical requirements, we are also interested in usability issues. That is, the proposed alarm should be relatively easy for the target user community to transition into use. To this end:

- (6) The proposed implementation must be practically feasible in terms of cost and additional equipment.
- (7) The alert must be safe, comfortable, and easy to use.
- (8) The target users must be willing to use the device.

4.1 Prototype Design

Based on our research and field investigation in the traffic operation scenario outlined above, we explored the use of two auditory and two haptic alerts [16]. The two auditory alarms, which varied in sound pattern and other tonal characteristics, were designed to emanate from the cruiser loudspeaker system. The haptic alerts varied in location of wear. The designs of the prototypes are outlined below.

Auditory Prototype Design. The State Police cruisers currently use three of ten preprogrammed siren tones on the SA314 series of Whelen box amplifier sirens, commonly referred to as *Wail*, *Yelp*, and *Piercer*. From a practical standpoint, it would be a relatively effortless and low-cost choice to activate one of the currently unused siren tones. For this reason, we chose two of the remaining seven signal tones as prototypes for the officer alerting mechanism. The first, *Pulsed Airhorn* consists of a repeating two pulse tone, which repeats about every second. The second, *Woop*, is a repeating single tone that increases in pitch over a period of about 250 ms. These two particular signals were selected for their distinguishability from the sirens currently in use and roughly evaluated on urgency by pitch and period. Other available signals had longer periods (lowering the perceived urgency), or were similar to *Piercer*, *Yelp*, and *Wail*, sirens already in use.

These proposed sirens also have several desirable characteristics consistent with our prototype requirements and communicated the appropriate semantics for a high priority alert. First, both sirens have varying tonal characteristics, which are important for alert discrimination and recognition. The human auditory system is much better at perceiving changes in sounds than pure tones [17]. In terms of sound intensity, it is suggested that the signal have a 10 to 30 dB increase over the ambient environmental noise with a maximum of 90 dB [20]. Our experience during the ride along showed that highway sound can reach levels around 80 dB, so our proposed signal should be 90 dB. The Whelen siren is capable of reaching this volume at close range.

Haptic Prototype Design. The haptic warning device designed for this system is a small wearable device that delivers a vibration signal when triggered. This trigger can be activated wirelessly from a computer and works in all weather conditions. The haptic device was engineered in the lab. The system diagram is shown in Fig. 3.

To achieve wireless communication, we use XBee wireless radio frequency modules with both a 300-foot and one-mile range. When a hazard is detected by the machine vision system, a serial command is sent to the transmitting XBee from the computer, which will then transmit a trigger signal to the receiving XBee. The receiving XBee is connected to an Arduino Fio, a smaller version of the Arduino microcontroller specifically designed for wireless applications. The Fio powers an

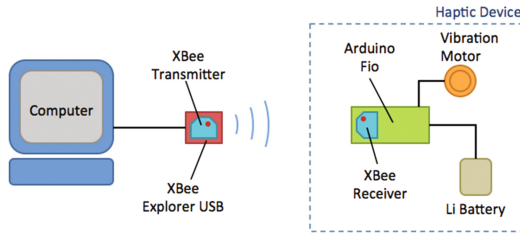


Fig. 3. Haptic system diagram.

eccentric rotating mass motor rotating at 12000 RPM to cause device vibration. These types of motors are similar in mechanics and intensity to those used in cellphones, game controllers and other vibrating devices. To power the motor, the system requires a small circuit—a transistor, resistor, and diode combination (not pictured in Fig. 3). The device is encapsulated in a custom-made case using a 3D printer (Fig. 4). The case features a small belt loop through which an elastic band can be threaded.

In experimentation, we were interested in testing this device on the wrist and on the waist. Ideally, a haptic device would be integrated into something that the officer already wears such as a watch, or belt. The wrist and waist were thus chosen to mimic this kind of integration and also for their sensitivity relative to other locations on the body. In both these locations, the motor was placed in contact with the skin.

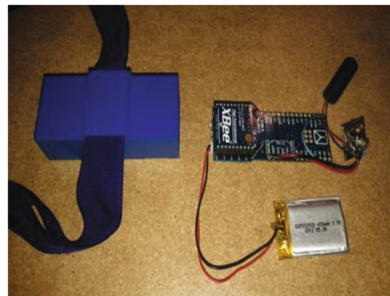


Fig. 4. Haptic device hardware, consisting of Arduino Fio, battery, and vibration motor and fabricated case.

5 User Evaluation

To assess the usability and effectiveness of the proposed prototypes, two studies were conducted: a human subjects experiment, and a focus group assessment with members of the user community—the State Police.

5.1 Human Subjects Experiment

In this experiment, we were interested in identifying which of the four prototypes, *Wrist Haptic*, *Waist Haptic*, *Pulsed Airhorn*, or *Woop*, would induce the fastest response, and which was most preferred by users. We set up a lab environment to replicate a high noise, low-light environment, and collected data on response times and subjective feedback in order to compare the effectiveness of each alert type with one another [16].

Design. The experiment was a four condition within-subjects experiment. That is, each participant interacted with each of the four alert types at the time of experimentation.

Independent and Dependent Variables. The experiment consisted of four randomized and counterbalanced trials. These trials were identical with the exception of the alert type used: *Wrist Haptic*, *Waist Haptic*, *Pulsed Airhorn* siren, or *Woop* siren. Thus, the *alert type* was the independent variable. During each session, we recorded the alert type, time at which the alert was triggered, and the time between the alert trigger and response to the alert (a key press). We also surveyed each participant on various aspects of the alarm to collect a subjective response. Thus, the dependent variables were *response time* and *subjective feedback*.

Participants. Forty volunteers (17 male) were recruited and screened to exclude those with known hearing impairments.

Apparatus. The study was conducted in a sound-proof room. To simulate the operating environment, recordings of ambient noise taken during the ride along were played over stereo speakers located on the right and left sides of the room. These speakers were connected through an amplifier box to a dedicated laptop, which controlled the audio playback. The decibel level of this playback varied between 70–77 dB. In our research and design phases, we concluded that highway noises may reach up to 80 dB and the optimal alarm decibel level might be 90 dB (a 10 dB increase over the max environment level). In our study, however, these levels were slightly reduced for hearing safety reasons. A third speaker was placed in front of the participant and connected to a second laptop to be used for the auditory signals. This second laptop was also connected to a wireless transmitter that could trigger the haptic alert. It ran software that controlled the type of alert, time at which it was triggered, and data logging over the course of the procedure. The room was illuminated by only a small lamp on the ceiling to replicate street lighting at night. An iPad 2 was provided to perform a secondary task.

Task. In runs in which the participant was outfitted with a haptic warning device, he or she was instructed to press a key on the laptop placed in front of him or her when the warning mechanism vibrated. In the two other sessions, the participant performed the same action in response to an auditory signal played from the speaker located in the front of the room. During all sessions, each participant was instructed to stand in the center of the room and play any of several games on an iPad to provide cognitive stimulation, as a proxy for cognitive demand an individual may experience outside of the alerting mechanism itself. The selection included the games, “Supermagical”,

“Angry Birds”, “Unblock Me”, “Candy Crush”, “Icomania”, “Jetpack”, “Blitz”, “Temple Run 2”, and “CollapseBlast”. The purpose of this secondary task was simply to distract the participant from focusing their attention on hearing or feeling the alert. This kind of sensitivity to a particular channel of perception would detract from the authenticity of the lab environment. We chose iPad games, a visual and kinetic task, to mimic simple, common police assignments such as writing citations. These particular games we chosen from popular games for iPad to ensure an engaging selection.

Procedure. Each session was preceded by a practice trigger in which the participant was given the opportunity to experience the alert in the upcoming session but not respond to it. Once the test session was started, playback of the ambient highway noise began, and the participant was directed to begin playing an iPad game of choice while standing facing the alarm speaker. The alarm signal was automatically triggered by the computer at a preselected time.

Each alert was triggered only once during a test session. To select alert trigger times, a sequence of forty times between thirty seconds and eight minutes were randomly selected, one for each participant, and a random permutation of these forty times was used for each type of alert. Thus, the average trigger time for each alert type across all experiments was identical.

The time at which a key press was detected in response to the alert was automatically recorded and the test session ended ten seconds after the alarm was triggered. Following each session, each participant was asked to complete a questionnaire to gather subjective information about his or her interaction with the warning signals.

5.2 Officer Assessment

Following the experiment, we conducted a focus group with members of the State Police to understand their perspectives on the proposed warning prototypes [16]. Findings from this are discussed in the next section.

6 Results and Discussion

Of the forty participants, ten were missing response time data for at least one of the four conditions due to test bed and subject errors. For example, in cases where an alert malfunctioned or the subject did not respond to the alert in the appropriate way, accurate readings were not taken.

6.1 Response Time

Response time is the key measure as the warning needs to induce a reaction within seconds. We ran a one-way repeated measures ANOVA using response time as the dependent variable and the alert type as the independent variable with 4 levels. There was a significant effect of alert type on response time ($F(3,87) = 27.5$, $p < .0001$) indicating that some alert types induced a significantly faster response than others.

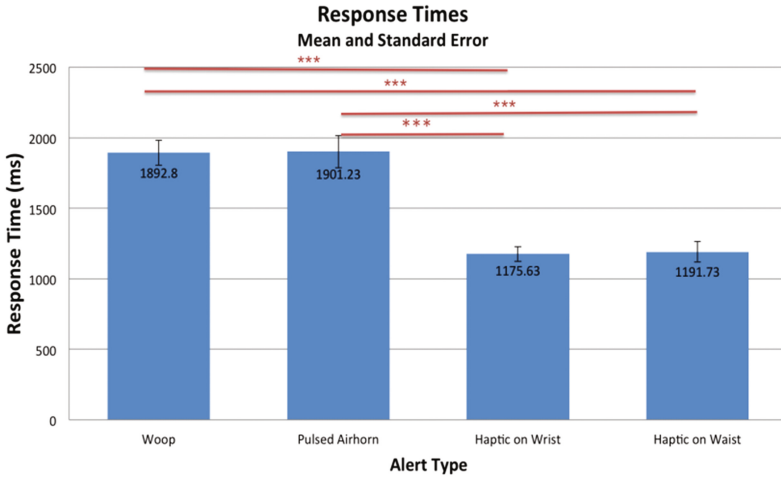


Fig. 5. Mean and standard error of response times in each condition. Red indicates significance: *** $p < .001$.

A post hoc Tukey's pairwise comparison revealed the significant differences between *Woop* and *Wrist Haptic* ($p < 0.001$), between *Woop* and *Waist Haptic* ($p < 0.001$), between *Pulsed Airhorn* and *Wrist Haptic* ($p < 0.001$), and between *Pulsed Airhorn* and *Waist Haptic* ($p < 0.001$). Other pairs showed no significant differences. These results indicate that the modality of warning had a significant effect on the response time. More specifically, responses to haptic signals were about 0.7 s faster than responses to the auditory signals. Moreover, considering that two of each signal modality were studied, the effect seems repeatable in experimentation. Figure 5 illustrates the mean and standard error of the four conditions.

6.2 Subjective Data

In terms of subjective data, study participants were asked to rate several features of the haptic and auditory alerts using a five point Likert scale. Specifically, subjects rated the intensity of the volume and pitch for the auditory alerts, the comfort of vibration, wear, and movement wearing the device for the haptic alerts, and detectability, signal urgency, warning appropriateness, and warning effectiveness for all four alerts. According to Wilcoxon matched pairs signed-rank test, there was no significant difference in *volume* ratings between the two auditory signals and no significant difference between the haptic alerts in *comfort of vibration* or *comfort of wear*. However, a Wilcoxon matched pairs signed-rank test showed a significant effect of the type of auditory signal on ratings of *pitch* ($W = 78$, $Z = 6.19$, $p < 0.005$, $r = 0.565$). Pitch was rated from "Too Low" to "Too High." The mean and standard error of the pitch ratings are plotted in Fig. 6. On average, subjects felt that the *Woop* siren was higher in pitch than the *Pulsed Airhorn* and tended to rate it closer to the "Too High" end of the scale. For each of the four prototypes, subjects were also asked to rank *detectability* on a scale

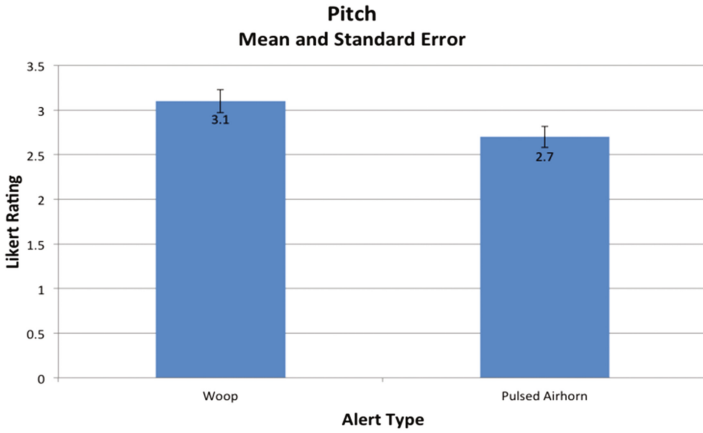


Fig. 6. Mean and standard error of pitch ratings in each auditory alert condition.

from a one, “Very Difficult to Detect” to five, “Very Easy to Detect.” A Friedman test revealed no significant difference between the four conditions. However, a significant effect was found of alert type on ratings of *urgency* ($X^2(3) = 33.945, p < 0.0001$). Urgency was rated from “Very Relaxed” to “Very Urgent.” The *Woop* signal was rated as significantly more urgent than the other three indicating that it would be easiest to detect (Fig. 7). A post-hoc test using Dunn’s Multiple Comparisons Test showed significant differences between *Woop* and *Pulsed Airhorn* ($p < 0.01$), between *Woop* and *Wrist Haptic* ($p < 0.001$), and between *Woop* and *Waist Haptic* ($p < 0.001$).

Finally, we found a significant effect of alert type on *effectiveness* rating ($X^2(3) = 21.514, p < 0.0001$) with significant differences between *Pulsed Airhorn* and

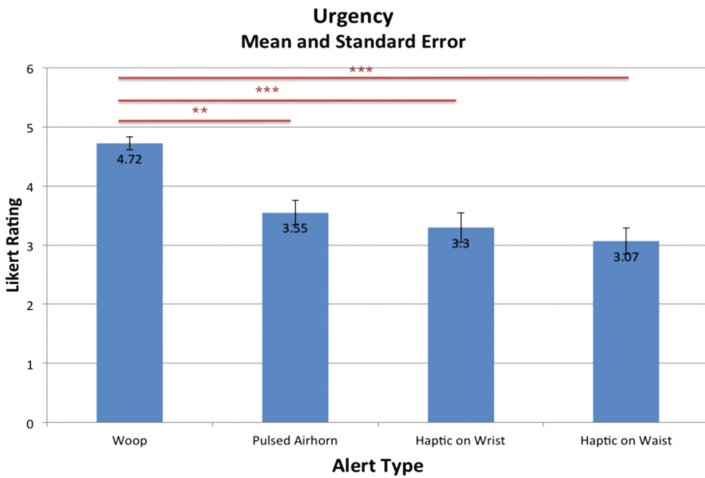


Fig. 7. Mean and standard error of urgency in each condition. Conditions that were significantly different are indicated in red: ** $p < .01$, *** $p < .001$.

Waist Haptic ($p < 0.05$), between *Woop* and *Wrist Haptic* ($p < 0.05$), and between *Woop* and *Waist Haptic* ($p < 0.01$). In these comparisons, the auditory sirens were rated higher than the haptic conditions as seen in Fig. 8.

In addition to these Likert scale ratings, the subjective survey concluded with a request for rankings on all four prototypes based on preference (“1” being the most preferred and “4” being the least preferred). A Friedman test here revealed a significant effect ($X^2(3) = 11.427$, $p < 0.01$). Dunn’s multiple comparisons test only showed a significant difference between *Pulsed Airhorn* and *Waist Haptic* ($p < 0.05$) in which *Pulsed Airhorn* was, on average, rated higher than the haptic signal located on the waist. Overall, *Pulsed Airhorn* was the most preferred. The results are summarized in Fig. 9.

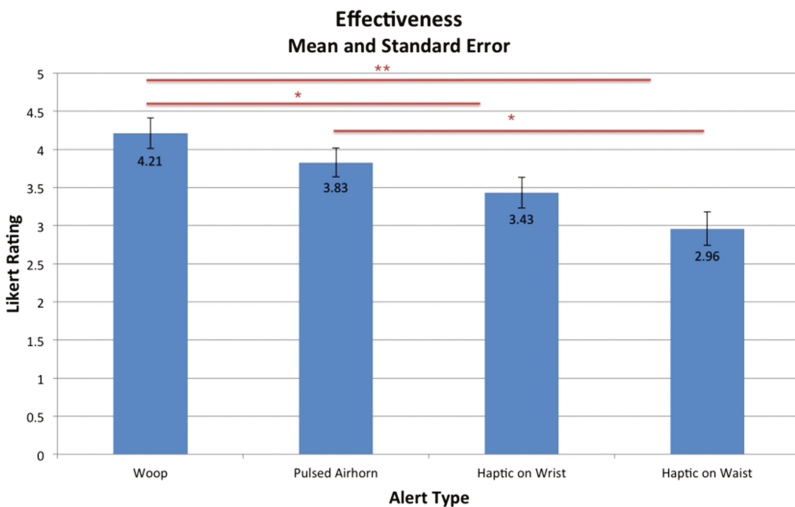


Fig. 8. Mean and standard error of effectiveness across conditions. (* $p < .05$, ** $p < .01$).

6.3 Comments

In general, comments varied in terms of whether subjects preferred the auditory or haptic signal. There was some general consensus, however, on various aspects of the individual prototypes.

For the haptic warning on the waist, the vibration was generally perceived as detectable and comfortable although many participants likened the vibration to a cell phone vibration or that of other similar devices. One participant stated, “The vibration frequency wasn’t ‘relaxed’ but seemed along the same ‘force’ as a hand held massager so doesn’t exactly bring emergency to mind.” Some even felt that the vibration was “ticklish.” It seems that because of its similarity to sensations we have naturally learned to associate with other devices, the haptic alert loses its novelty and hence its perceived urgency. Another common response to the haptic device was that it was at least mildly uncomfortable to wear, typical of an early prototype. This wearability issue should be

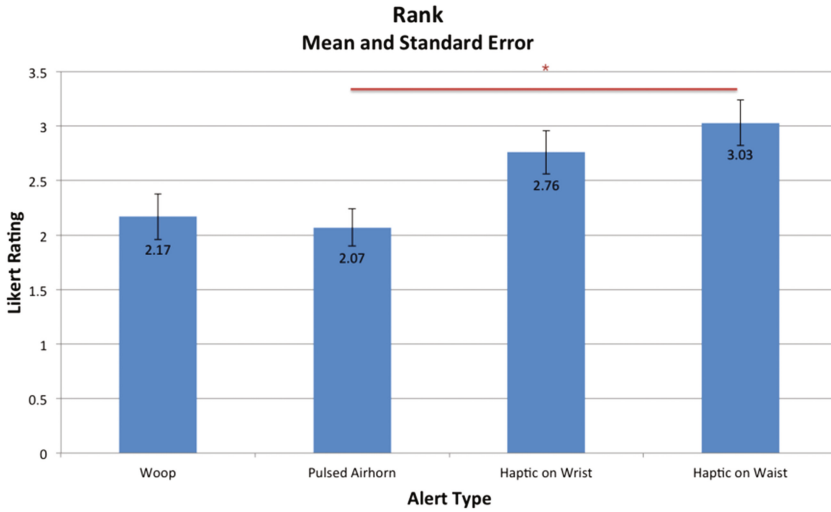


Fig. 9. Mean and standard error of rank rating in each condition. Red indicates significance: * $p < .05$.

easy to improve in further revisions. For example, the current device was designed with pointed corners, which could be rounded for better ergonomics and the hardware could be modified to be less bulky to wear. Eventually, the device can be integrated into an existing device or garment.

Some responses to the haptic device on the wrist were similar to those with the haptic on the waist in terms of the quality of the vibration and wear. It was described as “a little unwieldy” and “enough to signal/alert without stressful disturbance.” In general however, many subjects compared the device on the wrist to a watch, a location that felt more natural than the waist. One common sentiment was that the vibration on the wrist was “much more comfortable than on the stomach.” An important advantage of incorporating vibration into a device such as a watch would be that the vibrating mechanism would be much more likely to maintain contact with the skin.

In regards to the two auditory alarms, subjects tended to perceive the volume of both auditory alarms as “definitely audible” but also very close to the ambient noise. Although the sounds were controlled at 90 dB, about 10 dB higher than the ambient noise, it was common for subjects to observe that either auditory signal was, “loud by itself but not when the background noise was on.” In regards to the *Woop*, one subject commented, “Pitch was slightly on the high side but I feel that it stimulated an appropriate response,” while many responded to the *Pulsed Airhorn* with remarks such as, “Could be higher.”

In regard to the two sirens, comments included, “Catches my attention very well,” and, “Couldn’t have done a better job.” However, a common observation for both auditory signals was that they sounded similar to regular highway noises such as “truck horns” and “an actual siren.” If this concern proves to be an issue in the field, it could be mitigated by engineering new and more unique sounds for the operational environment.

6.4 Focus Group Findings

Following the conclusion of the user study, we conducted a focus group with four members of the State Police force to demonstrate the prototypes and gather feedback from experts in the field. The experience of the four individuals ranged from 17 to 31 years on the State Police force.

In response to the haptic signal, an officer wearing the device during the demonstration commented that the vibration “caught my attention right away” and all four agreed that the intensity was appropriately strong and different from that of a cell phone vibration. There were, however, varying opinions on the optimal location of wear. As a watch, some felt that it would be best in terms of maintaining the effectiveness of the device, but that most officers don’t wear watches and that it would easily be forgotten. Another suggestion was to instead, integrate the vibration into the duty belt because, “you are always going to put it on.” However, there were concerns as to how easily the vibration would be felt through layers of clothing or when standing or sitting in different positions.

The officers also suggested putting the device in a pocket and/or modifying uniforms to have holes where the motor could be placed in contact with the skin. If integrated into the clothing, there would be a need for multiple devices for each officer – one for each uniform. Another idea was to wear the device as a necklace and the other individuals seemed to agree that this was a viable option. When asked whether officers would be inclined to wear the device on a regular basis, the general consensus was positive. In discussion of the haptic device, we also learned that in terms of battery life, the haptic device would need to run for up to 16 h (the length of a double shift).

In response to the auditory alarms, all four officers agreed that they preferred the *Woop* over the *Pulsed Airhorn*. The *Pulsed Airhorn* was “too similar to the air horn we already use.” With the auditory alarm, there was also the concern that it would go off in a situation in which an officer would not want to bring attention to himself or herself (for example when watching a scene before going in). However, one of the officers acknowledged that the siren would not go off unless the emergency lights were on, based on the programming of the cruisers, and then the others seemed to agree that this was acceptable. The officers also agreed that in all cases, the warning should automatically turn off after a ten-second timeout.

Next, when asked if a multimodal warning incorporating the *Woop* signal and haptic device would be useful, the answer was a resounding yes. The auditory signal would “always be there” since it would be a part of the cruiser hardware and the haptic signal would be supplemental.

Overall, the officers were enthusiastic about the prototypes. One of the individuals, serving as director of fleet operations and responsible for equipment, stated, “It’s a great tool, I really do think,” and concluded saying that, “If we can absorb that cost, it’s a no-brainer.”

7 Future Work

There are some features that we did not implement in the current prototypes that are worth investigating in future work. The existing prototypes could easily be integrated with each other or other modes of warning to create a multimodal alerting mechanism. With such a warning, it would be worthwhile to study whether a warning that uses multiple modalities can improve response time over a single modality.

In terms of modifying the existing prototypes presented in this paper, several changes could be made to improve feedback from the usability studies. First, we did not investigate the design of a new sound with the desirable qualities of signal urgency and conspicuity for these environments. It is possible that a unique sound tailored to the environment could produce superior response times.

Second, the haptic device could be upgraded in two ways. First, it currently delivers a continuous vibration but may benefit from a modification in intensity of the signal or in a change in vibration pattern. In addition, a more comfortable device design could improve user acceptance. Ideally, the haptic signal could be integrated into a device that the user already wears on a regular basis, such as a belt and based on the conversation with the state police, there is also work to be done in pinpointing the best method and location of wear.

8 Conclusion

Based on fieldwork and background research, and in close collaboration with state police, four alert prototypes were designed and evaluated for use in a high noise, low-light environment such as a dimly-lit highway shoulder. Two of these alerts were auditory sirens and two were haptic vibrations, one placed at the wrist and one at the waist. Haptic vibrations, which we hypothesized would be more salient in a loud and visually stimulating environment, produced statistically significantly faster responses than the auditory alerts. However, there were no statistical differences between the two haptic and the two auditory alerts, suggesting robustness to the specific alarm type.

Subjectively, the subjects had a slight preference for the auditory alerts and perceived them as significantly more urgent than the haptic alerts. However, both the subjects and state police officers responded positively to the haptic alerts overall. The discrepancy in the findings, objective data leaning towards haptic and subjective data leaning toward auditory, could be attributed to cognitive fluency with auditory sirens over haptic alerts. Subjectively, participants favor auditory alerts in this scenario because they are more familiar and feel easier to process even if the data shows otherwise. In practice, while the auditory alerts offer more permanence and durability, the haptic alerts are a more novel and thus more conspicuous stimulus in the operation environment. It was proposed by members of both groups that a multimodal alert using both signal types could be highly effective.

In a traffic environment, most existing safety technologies focus on passive danger prevention rather than active warning which can incorporate autonomous safety technology as proposed in this work. This research has shown that once alerted through a haptic device, a person in this low-light setting gains, on average, an additional 0.7 s

in response time as compared to an auditory alert, which could mean the difference between life and death. However, the fidelity and reliability of the overall system, i.e., the sensors and algorithms that detect a possible oncoming threat, will determine whether these alerting schemes are successful. If such a system experiences many false positives, users may become frustrated and learn to ignore the system. Highlighting the importance of systems engineering, the overall testing of the integrated system, which is still pending, will determine the ultimate success of these alerts.

Acknowledgments. NSF (NSF Grant#1136996 awarded to Computing Research Association for the CI Fellows Project) and the National Institute of Justice partially sponsored this effort. The Massachusetts State Police and the Cambridge Police Department were instrumental in this effort, and we would also like to acknowledge Seth Teller, Berthold Horn, and Brian Wu of the MIT Computer Science and Artificial Intelligence Laboratory for their collaboration on this project.

References

1. Ashton, R.J. Solutions for safer traffic stops. *Police Chief* (2004)
2. Belz, S.M., Robinson, G.S., Casali, J.G.: A new class of auditory warning signals for complex systems: auditory icons. *Hum. Factors* **41**(4), 608–618 (1999)
3. Burt, J.L., Bartolome, D.S., Burdette, D.W., Comstock, J.R.: A psychophysiological evaluation of the perceived urgency of auditory warning signals. *Ergonomics* **38**(11), 2327–2340 (1995)
4. Cergol, G.: On-Duty Nassau County Patrol Officer Struck, Killed by Car on LIE. NBC, New York (2012)
5. Cummings, M.L., Donmez, B., Graham, H.D.: Assessing the Impact of Haptic Peripheral Displays for UAV Operators (2008)
6. Daoud, S.O., Tashima, H.N.: 2012 Annual Report of the California DUI Management Information (2012)
7. Edworthy, J., Stanton, N.: A user-centered approach to the design and evaluation of auditory warning signals. *Ergonomics* **38**(11), 2262–2280 (1995)
8. Feathers, T.: Trooper injured when alleged drunk driver crashes into cruiser in Saugus. *boston.com* (2013)
9. Federal Highway Administration. Boosting Roadway Safety with Rumble Strips (2002)
10. Haas, E.C., Schmidt, J.: Auditory icons as warning and advisory signals in the US Army Battlefield combat identification system. *HFES* **39**, 999–1003 (1995)
11. International Safety Equipment Association. American National Standard for High-Visibility Safety Apparel and Headwear (2010)
12. Joint Transport Research Centre of the OEDC and International Transport Forum. Towards Zero: Ambitious Road Safety Targets and the Safe System Approach Summary Document (2008)
13. Karraker, J.: Detecting, Tracking, and Warning of Traffic Threats to Police Stopped Along the Roadside. M.Eng. Thesis, MIT EECS, Cambridge, MA (2013)
14. Michigan Department of Transportation. New Work Zone Sign: Give ‘em a Brake Safety Coalition Warns Motorists to Pay Close Attention (2006)
15. Patterson, R.D., Mayfield, T.F.: Auditory warning sounds in the work environment. *Phil. Trans. R. Soc. Lond.* **327**, 485–492 (1990)
16. Powale, P.: Design of an Alerting Device for Roadside Personnel, June 2013

17. Provins, K.A., Morton, R.: Tactile discrimination and skin temperature. *J. Appl. Physiol.* **15** (1), 155–160 (1960)
18. Reisberg, D.: *Cognition*. Norton & Company Inc., New York (2007)
19. Richard, C.M., Brown, J.L., McCallum, M.: Crash warning system interfaces: human factors insights and lessons learned (No. HS-810 697) (2007)
20. Scott, J.J., Gray, R.: A comparison of tactile, visual, and auditory warnings for rear-end collision prevention in simulated driving. *Hum. Factors* **50**(2), 264–275 (2008)
21. Spence, C., Ho, C.: Tactile and multisensory spatial warning signals for drivers. *IEEE Trans. Haptics* **1**(2), 121–129 (2008)
22. The Associated Press. Officer escapes injury in traffic stop accident. *The Record* (2012)
23. U.S. Department of Justice and Federal Bureau of Investigation, C.J.I.S.D. Law Enforcement Officers Killed and Assaulted (Table 61: Law Enforcement Officers Accidentally Killed; Circumstance at Scene of Incident, 2000–2009) 2009
24. U.S. Fire Administration. Emergency Vehicle Visibility and Conspicuity Study (2009)
25. Wogalter, M.S. (ed.): *Handbook of Warnings*. Lawrence Erlbaum Associates, Mahwah (2006)
26. Wood, N.E.: Shoulder rumble strips: a method to alert “Drifting” drivers. In: *Proceedings of the 73rd Annual Meeting of the Transportation Research Board* (1994)
27. Wu, B.: *A Controllable Laser Projector for Diverting Traffic*. M.Eng. Thesis, MIT EECS, Cambridge, MA (2013)
28. National Campaign Launches Effort Educating Drivers to ‘Move Over’ and Protect Officers on Roadways. *Move Over, America* (2007)

Clustering of in-Vehicle User Decision-Making Characteristics Based on Density Peak

Qing Xue, Qian Zhang^(✉), Xuan Han, and Jia Hao

School of Mechanical Engineering, Beijing Institute of Technology,
100081 Beijing, People's Republic of China
{xueqing, haojia632}@bit.edu.cn,
Zhangqian_bit@126.com, fleurdellys37hx@gmail.com

Abstract. In this paper, we designed the simulated combat experiment to obtain the decision of the participants. Combining with the characteristics of the decision - making in the combat procedure and combat task, the fuzzy recognition model was established to obtain the model user characteristic matrix. The decision-making characteristics clustering analysis of density peak is the foundation for the design of adaptive in-vehicle user interface based on user decision-making characteristics.

Keywords: Decision-making characteristics · In-vehicle user · Cluster analysis

1 Introduction

In the process of interacting with the vehicle interface and completing the operation task, the display content and mode of the interface play a very important role in the decision-making of the operator's behavior. In order to provide more helpful interface information and interactive mode for the operator's characteristics, the user can efficiently carry out the human-computer interaction, and the vehicle-based adaptive interface is the hotspot of the current research. Adaptive user interface(AUI) can be automatically adjusted according to the user's interactive behavior, thereby changing the interface information display and content, such as changing its layout, structure, style, display content and other elements to meet specific user requirements in a particular environment [1, 2]. AUI can predict user behavior, reduce the burden of human-computer interaction, improve the efficiency of interaction tasks, can adapt to the user, the environment or changing needs [3]. Although many scholars are working on the new vehicle interface adaptive mechanism, few researches on engineering vehicles and military vehicles have been done, especially, the user characteristics modeling are also lack of research in these field.

Without sufficient consideration of user characteristics, the lack of adaptive mechanisms for user characteristics is one of the causes of operator error in human-computer interaction. User modeling is an important part of AUI research. User model includes adaptive information, which can summarize and classify users. (1) In application-based modeling, Dieterich H [4] divides user characteristics into application-related computer experience knowledge and application-independent user

preferences and learning abilities. (2) In the aspect of cognitive-based modeling, Zuxiang Zhu [6] introduces the user's attention strategies, words and spatial abilities into the user model. (3) In the process of learning based on the user, Jingyun Cheng [7] select user thinking, learning in the understanding, memory and other characteristics of modeling; (4) Based on the number of users, Dieterich H et al. [4] Zhiwei Guan [5] and Xiao Li [8] model individual and user groups, general and special users, respectively.

The study of user classification includes classification criteria and classification methods. (1) In the study of user classification criteria, Dieterich H [4] Anthony F et al. [9] and Schiaffino S et al. [17] classify users both qualitatively and quantitatively. Zhiwei Guan [5] and Anthony F [9] studied the frequency of use, purpose, type of learning and proficiency of users, and develop the standards for classification. (2) In the study of user classification methods, Hongmei Ge et al. [10] carried out two naive Bayesian classifications according to the user's interest vector in all time slices, and got the user classification in the whole time period. LingLuo [11] used the fuzzy c-Means clustering algorithm to cluster the power users according to the load curve of power users. Wei Zheng [12] proposed a Web user clustering method by extracting the user's access behavior and getting the similarity. And use these clusters as the previous empirical data for the artificial neural network, improve the efficiency of user classification.

Decision-making style is a decision-making habits, but also for accepting or reflecting the decision-making tasks personal characteristics [13]. For the type of decision-making style, different scholars according to different standards were classified. Henderson and Nut [14] divide the decision style into Analytic and Heuristic according to the rational way. Driver [15] divides decision styles into five types according to the amount of information used. Scott and Bruce [16] divided decision making into five different styles.

Based on the human-computer interaction task, this paper established a combat simulation experiment based on abstract task content and information. Design experimental tasks and human-computer interaction interface with LabVIEW. The related indexes in the task performance of participants were obtained. The user characteristic matrix in the form of fuzzy recognition model was obtained based on the characteristics of the decision process presented by the participants in the experiment. Cluster based on density peak and the analysis result is the foundation for the design of adaptive in-vehicle UI based on user decision-making characteristics.

2 Method

This experiment is simulated combat experiment, the participants complete the given task as a real user in battlefield.

The experiment was carried out to obtain the decision of the participants in 5 kinds of combat situations.

In order to reduce the participants' unfamiliarity with the effect of the vehicle interface on the experimental results, the participants need to repeat the experiment several times.

The experiment is divided into stage 1 and stage 2.

Stage 1 is user modeling.

Stage 2 is user clustering.

2.1 Experiment Design

Interface Design

Vehicle interface is the channel between user and the vehicle to transmit information, including the display and controller. The display design is a vital part of the overall design of the vehicle interface, which directly affects the overall combat effectiveness. In this experiment, the vehicle interface display is designed to provide the user with full consideration of the operating characteristics, allowing vehicles and users to adapt to each other and improve the efficiency of interaction.

In this experiment, we designed the vehicle interface display information and layout.

Information

Combined with specific combat situation, the information provided to the participants included 4 parts.

(1) Map

Including one's own position with the enemy's information, the number of the enemy, the direction of the front.

(2) Description

Description describes the current combat situation, the experiment contains 5 kinds of combat situations.

Situation 1: The enemy is located outside the scope of the attack.

Situation 2: The enemy is located within the scope of the attack.

Emergency of Situation 2: When the enemy is within the attack range, a new enemy is detected.

Situation 3: the enemy is located within the scope of the attack and the enemy has launched an attack.

Emergency of Situation 3: The enemy is located within the attack range and the enemy has launched an attack, a new enemy is detected.

(3) Task information

Including the degree of threat of the enemy, the relative position (distance and height), target types and attack plans.

The relative position includes distance and height difference; the target type is divided into helicopters, tanks and fighters; attack plan contains different enemies equipped with the corresponding attack methods, namely shells, missiles and shells together.

After abstracting and simplifying the combat information, we designed the simulated combat task. The information of the task in 5 cases is shown in Table 1, in which

Table 1. Task information in 5 situations

		Situation 1	Situation 2	Emergency of situation 2	Situation 3	Emergency of situation 3
Threat degree	A	0.929	0.803	0.695	0.004	0.003
	B	0.004	0.018	0.014	0.923	0.979
	C	0.975	0.943	0.948	0.899	0.856
	D			0.629		0.798
Relative position (Distance–Height)	A	30-50	14-20	14-20	14-20	14-20
	B	35-5	8-5	8-5	7.1-100	7.1-100
	C	45-2000	25-1500	25-1500	18.2-2500	18.2-2500
	D			3.1-15		3.1-5
Target type	A	H	H	H	T	T
	B	T	T	T	H	H
	C	F	F	F	F	F
	D			T		T
Attack plan	A	M	M	M	S	S
	B	S	S	S	M	M
	C	M	M & S	M	M	M
	D			S		S

Notes:*H* represents *helicopter*; *T* represents *tank*; *F* represents *fighter*
M represents *missile*; *S* represents *shell*

the threat degree is calculated from the relative position and the target type, and the attack plan is given according to the specific situation.

(4) Feedback

Feedback the operation of the participants.

Layout

According to the requirements of the experiment, the display should include the following five parts: the map area, task information area, description area, feedback information area and control button area. Meng Wang [18] studied the identify performance on the display location information recognition, and the result is 1 (upper left area) > 2 (lower left area) > 3/3' (upper right and lower right areas), as shown in Fig. 1. The user has the highest recognition identify for the information displayed by 1 (upper left area), so that the map area presenting the complex information is arranged at 1 position; Wherein the user has a faster reaction speed and a higher correct rate for the information displayed by the 2 (lower left area), and is suitable for displaying important prompt information, so that the description is arranged at the 2 position; The user has a high accuracy rate for the information displayed at the 3' (lower right area), thus arranging the feedback at the 3' position; The user has a faster response speed to the information displayed by 3 (upper right area), so that the task information area is arranged at the 3 position.

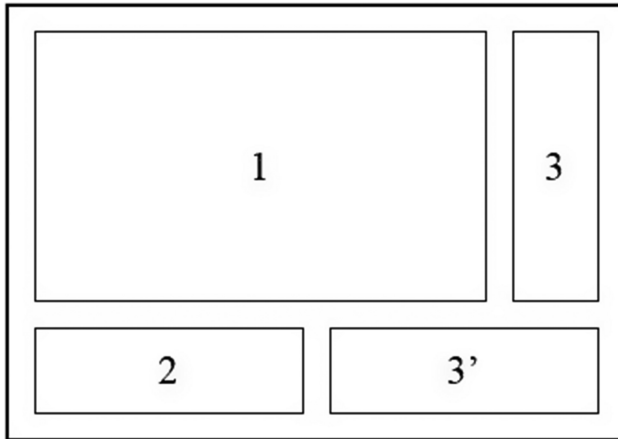


Fig. 1. The sorting of identifying performance of display parts

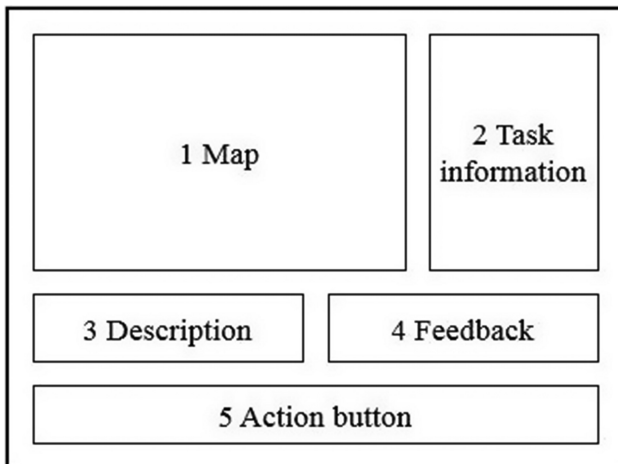


Fig. 2. Vehicle man - machine interface layout

After taking into account the above information and the actual design of the display, we determine the final layout of the display shown in Fig. 2.

Task Procedure

In order to reduce the risk and the difference between operators in manipulating complex controllers, the participants were asked to use a mouse to perform the experiment tasks.

In this experiment, participants need to follow the instructions, according to the battle map and task information and make decisions (attack, defend, escape) in the control button area, the decision will be given feedback. Participants need to make

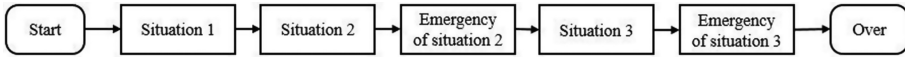


Fig. 3. Experimental flow chart

decisions in 5 situations in turn, and the experimental flow of a simulated combat mission is shown in Fig. 3.

Data Collection

In this experiment, LabVIEW was as the software environment background. The program recorded the decision under different experimental conditions and the number of experiments.

A total of 41 undergraduates, graduate students and doctoral students from Beijing Institute of Technology (including 29 males and 12 females) participated in the experiment. They are between the ages of 20–29 years of age, visual acuity and corrected visual acuity were normal, and has not participated in the test of such experiments. Taking into account the effect of fatigue on the experimental results, the experiment time should not exceed 45 min.

2.2 Stage 1: Modeling of User Decision-Making Characteristics

Select 10 participants, and 10 combat tasks were repeated. Combining with the combat process and combat task, a fuzzy recognition model of user characteristics are established, which can identify the interaction characteristics of the user through the interaction behavior. Analyze the convergence of the user model as the number of user experiments increases.

User Model

Combining with the experimental background of the armored vehicle combat and the behavioral characteristics exhibited by the participants, this study proposed the characteristics of “conservative”, “calm” and “risk” to measure the user’s decision style. Each of the user’s decisions are reflected in these three characteristics.

In this study, the fuzzy comprehensive evaluation method is used to identify the user’s decision making [19]. This method is based on the evaluation results of the single factor related to the evaluation object. Finally, the evaluation results are obtained by using the weight factor of each factor.

$X = \{x_1, x_2, \dots, x_n\}$ is a set of evaluation factors, and $Y = \{y_1, y_2, \dots, y_m\}$ is a set of evaluation indexes. Each user is determined a fuzzy relation R from X to Y , R is the evaluation matrix of the user, in which the various factors of r_{ij} in R mean evaluation object factor of x_i which belongs to the y_j membership of evaluation grade ($i = 1, 2, \dots, n$; $j = 1, 2, \dots, m$).

Λ is a fuzzy subset of X , indicating that the degree of importance of the evaluation factors, e.g. weight, $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$, where $\lambda_i \geq 0$ and $\sum \lambda_i = 1$.

C means user evaluation, $C = \Lambda \times R$.

In this paper, $X = \{x_1, x_2, x_3, x_4, x_5\}$, $Y = \{y_1, y_2, y_3\}$.

The $x_1, x_2, x_3, x_4,$ and x_5 respectively indicate the decision in the situation 1, situation 2, emergency of situation 2, situation 3 and emergency of situation 3. y_1, y_2, y_3 is characteristic value of risk, calm, and conservative. A fuzzy relationship from X to Y is as follows:

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \end{bmatrix}$$

In this experiment, five situations have the same influence on user behavior, so $\Lambda = [0.2, 0.2, 0.2, 0.2, 0.2]$ is taken.

The evaluation object factor x_i belongs to the y_j membership of evaluation characteristic which is determined by the membership function [20], getting the membership function of the user under different decision by the expert evaluation method [21, 22], adopted by the linear membership function distribution [23, 24], the membership

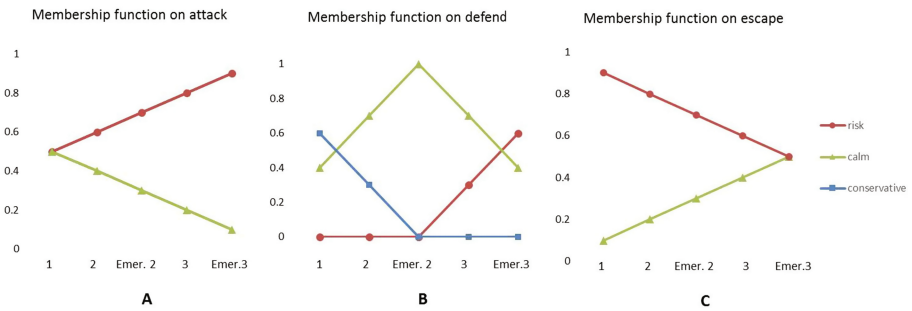


Fig. 4. The membership function of user characteristics under different behaviors

function of user characteristics under different behavior is shown in Fig. 4:

Result and Analysis

According to the user model in 2.2.1, we obtain the user evaluation matrix of 10 participants in stage 1. The changes of user characteristics in 10 experiments and the corresponding number of participants are shown in Table 2.

For example, in the case 1, the evaluation result of a user is always [0.42,0.36,0.22], the decision-making characteristic is stable and the risk characteristic is obvious; The evaluation results of a user in Case 3 are shown in Table 3, and the results of the evaluation have been changed in the first 6 trials. From the 7th trial, the results of evaluation tend to be stable, and the calm characteristics of this user are obvious. We argued that the user model stabilized after the 7th time of trial.

According to the current results, we only make a general analysis of the user’s decision-making characteristics, the conclusions are not necessarily applicable, we will further analysis user characteristics in stage2.

Table 2. User characteristics of the changes in stage1

	Characteristic	Number/percentage
Case 1	There were no changes in 10 experiments.	2/20%
Case 2	Changes in the pre-period, no more changes after 6 experiments.	5/50%
Case 3	The characteristic has been changing and tends to be stable after 7 experiments.	3/30%

Table 3. User evaluation in stage1

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10
0.28	0.32	0.12	0.24	0.12	0.24	0.42	0.42	0.42	0.40
0.52	0.44	0.58	0.46	0.44	0.52	0.40	0.52	0.52	0.48
0.20	0.24	0.30	0.30	0.44	0.24	0.18	0.06	0.06	0.12

2.3 Stage 2: Clustering of User Decision-Making Characteristics

The other 31 participants performed the trial for 7 times. According to the comprehensive evaluation of the user and the corresponding three evaluation indicators, the corresponding points in the 3-dimensional space are obtained. Cluster analysis based on density peak, and the decision-making characteristics of users in each cluster are analyzed to determine the decision-making style.

Result and Cluster Methods

At present, clustering methods are based on partitioning method, hierarchical method, density-based method and so on. In this paper, a method of fast search and find of density peaks (referred to as DPC) introduced by Alex Rodríguez and Alessandro Laio is used for clustering. The DPC algorithm is based on the idea that cluster centers are defined by having a higher density than their neighbors or at a relatively large distance from points with higher densities.

The clustering method defines two variables for each data point: its local density ρ_i and its distance δ_i from points of higher density. Both these quantities depend only on the distances between data points.

The local density ρ_i is defined as

$$\rho_i = \sum \chi(d_{ij} - d_c)$$

where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise, and d_c is a cutoff distance. ρ_i is equivalent to the number of dots whose distance to point i is less than d_c . As a rule of thumb and the number of data points in this experiment, we can choose d_c so that the average number of neighbors is around 5% to 8% of the total number of points in the data set.

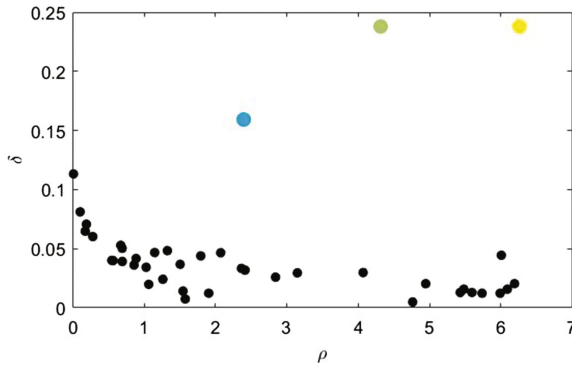


Fig. 5. Decision diagram (Color figure online)

δ_i is measured by computing the minimum distance between the point i and any other point with higher density:

$$\delta_i = \min_{j: \rho_j > \rho_i} (d_{ij})$$

The points with higher ρ_i and δ_i are considered to be clustering centers, and the remaining points are assigned to the clusters with higher density nearest to the point, and the clustering results are obtained.

Cluster Analysis

According to the calculation method in 2.3.1, calculate the ρ_i and δ_i of each data point, and the decision diagram is shown in Fig. 5.

It can be concluded from the Fig. 5 that the decision styles presented in this experiment can be divided into three categories, the cluster centers are three points of yellow (1), green (2) and blue (3). The number and percentage of users per cluster are shown in Table 4.

Table 4. Clustering results

Cluster	Number/percentage
1	19/46.34%
2	14/34.15%
3	8/19.51%

Results of cluster analysis are shown in Fig. 6A–D respectively. In the 3-dimensional space composed of three characteristics, the polygons formed by each cluster are not coincident. Users within each cluster have some common features that can be described by the users' 3 decision characteristics.

Figure 6A is a representation of the clustering of user characteristics in 3-dimensional space, and Fig. 6B–D are representations of two of the three properties in two dimensions respectively.

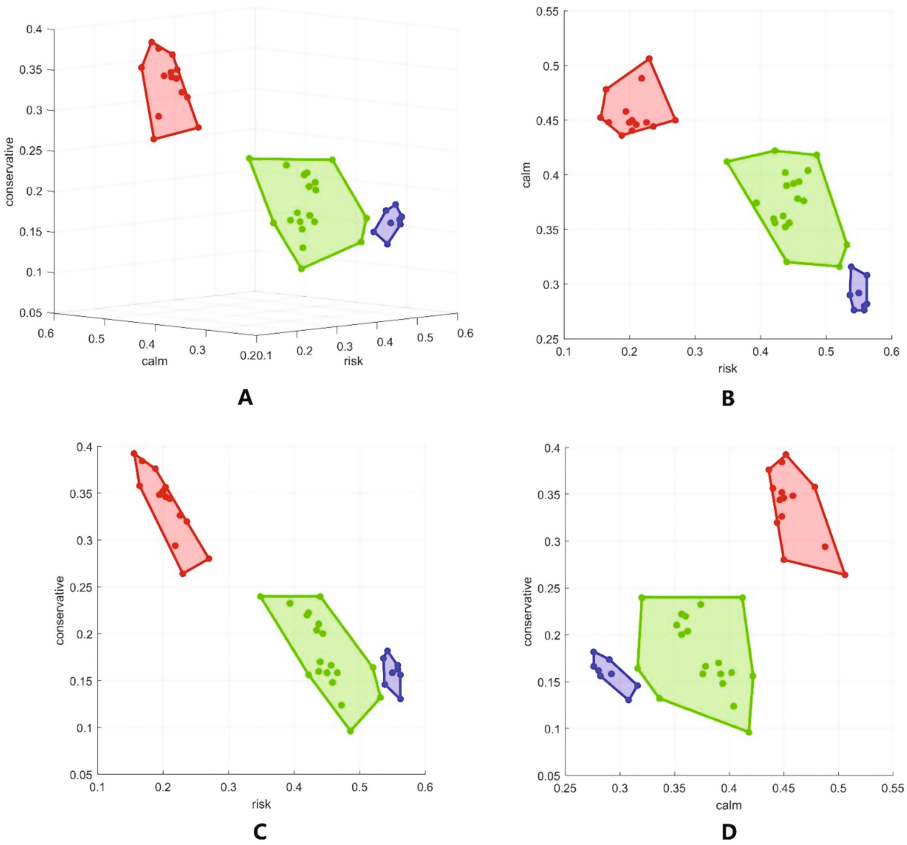


Fig. 6. The clustering results of user decision characteristics Notes: “●”represents cluster 1, “●”represents cluster 2, “●”represents cluster 3 (Color figure online)

Cluster 1 includes the largest number of points as well as the largest range, followed by cluster 2, and cluster 3 includes the smallest number of points as well as the smallest range. According to Fig. 6B and C, compared with cluster 2, the “risk” values of cluster 1 and cluster 3 are higher, and there are certain differences between cluster 1 and cluster 3. According to Fig. 6B and D, the “clam” values of cluster 2 are on high level, and the “clam” values of cluster 1 are on middle level, and the “clam” values of cluster 2 are on low level; besides, the “clam” values of the three clusters are continuously distributed, without significant differences between two adjacent clusters. It can be seen from Fig. 6C and D that the “conservative” values of cluster 2 are on high level, and the “conservative” values of cluster 1 and cluster 3 are on middle level, but the range of the “conservative” values of cluster 1 is higher than that of cluster 3.

3 Discussion

3.1 Stage 1

According to the user model and the experimental results in 2.2, we summarize all the user categories into 5 types, divided into 2 categories.

Category 1: The users with any decision-making character much higher than the other two characteristics in the evaluation result are the users of the characteristic, i.e., the risk type users, the clam type users or the conservative type users.

Category 2: The users with relatively high values of two adjacent decision-making characteristics but not an apparently higher one, i.e., the users with relatively high risk characteristic and clam characteristic as well as with relatively high clam characteristic and conservative characteristic in this experiment, are defined as partial risk users and partial conservative users, respectively.

3.2 Stage 2

According to Table 4, cluster 1 includes the largest number of participants, accounting for about half of the total number. Most of the “risk” values of users in cluster 1 are within the range of 0.4–0.5, accounting for about half of the whole proportion, and some users are lower than 50%. Most of the “clam” values of users in cluster 1 are within the range of 0.35–0.4, with little difference with the range of “risk” values. Comparing with the ranges of “risk” and “clam” values, “conservative” values have a wider range of 0.1–0.25, as the minimum values of the three characteristics. The users of cluster 1 have moderate risks and clams and low conservatives during decision-making, and they are partial-risk users.

Most of the “clam” values of users in cluster 2 are within the range of 0.43–0.5, accounting for about half of the whole proportion, and most users are lower than 50%. Most of the “conservative” values are within the range of 0.28–0.38, with larger range than the “clam” values. The “risk” values are the minimum ones in the three characteristics, within the range of 0.15–0.25. Although there is no significant difference between the “clam” values and the “conservative” values of the users in cluster 2, all the “clam” values are larger than the “conservative” values; therefore, the users in cluster 2 are conservative type users.

Cluster 3 includes the least participants, less than 20% of the total number. All the “risk” values in cluster 3 are about 0.55, more than half of the whole proportion. Both the “clam” values and the “conservative” values are relatively small, in which the “clam” values are larger. The “risk” values of users in cluster 3 are apparently the largest; with tendency of higher risks, the users in cluster 3 are risk type users.

Through the analysis of cluster 1, cluster 2 and cluster 3, we find that the user decision-making style is more risk and calm, and less conservative. The reason may be that it is difficult for the participants to have a more realistic experience of the urgency and danger in the experiment. So more decisions are attack and defense, less escape. Although there is a gap between the experimental situation in the experiment and the

situation in the real battlefield, but the simulated combat task has been able to achieve the user decision-making characteristic.

4 Conclusion

The main purpose of this paper is to classify the user based on the decision-making characteristics for the design of in-vehicle AUI. The paper is also the foundation for the future research of AUI design in special vehicle application.

In summary, the clustering based on density peak could be a practical way to research the in-vehicle user decision-making characteristics. In the 3-dimensions of “risk”, “clam” and “conservative”, the users of each cluster have different characteristics. The results of user decision-making cluster analysis are as follows.

- (1) According to decision-making style, users are divided into three categories, partial risk-type users is about half, and the rest for the users of calm and users of risk.
- (2) The partial-risk users show high risk and moderate calm as well as less conservative in decision - making. Users of calm show slightly greater calm and moderate conservative in decision making; Users of risk show greater risk and less calm and conservative in decision making.

Acknowledgements. The authors would like to thank the anonymous reviewers for their valuable comments and thank the strong support provided by National Natural Science Foundation of China (NSFC 51505032) and Beijing Natural Science Foundation (BJNSF 3172028).

References

1. Ge, L., Wang, Y.: Adaptive interface of computer - a new idea of computer interface design. *Chin. J. Ergon.* **3**, 50–52 (1996)
2. Van Velsen, L., Thea, V.D.G., Klaassen, R., Steehouder, M.: User centered evaluation of adaptive and adaptable systems: a literature review. *Knowl. Eng. Rev.* **23**(3), 261–281 (2008)
3. Gajos, K.Z., Czerwinski, M., Tan, D.S., Weld, D.S.: Exploring the design space for adaptive graphical user interfaces. In: Working Conference on Advanced Visual Interfaces, vol. 28 (4), pp. 183–191 (2006)
4. Dieterich, H., Malinowski, U., Kuhme, T., Schneider-Hufschmid, M.: State of the Art in Adaptive User Interfaces, March 2009. www.cc.gatech.edu/computing/classes/cs8133d94fall/ps-files/Siemens.ps.Z
5. Guan, Z.: Intelligent human - computer interaction oriented to user’s intention. Institute of Software, Chinese Academy of Sciences (2000)
6. Zhu, Z.: Engineering Psychology Course. People’s Education Press (2003)
7. Cheng, J., Ni, Y.: Human Machine Interface Design and Development Tools. Electronic Industry Press (1994)
8. Li, X.: Research on Adaptive Human Computer Interface. Southwest China Normal University (2004)

9. Norcio, A.F., Stanley, J.: Adaptive human-computer interfaces: a literature survey and perspective. *IEEE Trans. Syst. Man Cybern.* **19**(2), 399–408 (1989)
10. Ge, H.M., He, Y.X., Chen, Q., Xu, C.: Micro blogging users classification method based on the time slice. *J. Chin. Comput. Syst.* **34**(11), 2441–2445 (2013)
11. Luo, L.: Research on categorized time-of-use power price based on fuzzy C-means clustering. Shan Dong University (2013)
12. Zheng, W.: The Investigation of algorithm that Fuzzy & BP Neural Network for Classifying the Web Users. Zhejiang University of Technology (2012)
13. Driverd, M.J., Brousseau, K.R., Hunsakerh, P.L.: The Dynamic Decision Maker: Five Decision styles for Executive and Business Success. Jossey-Bass, San Francisco (1993)
14. Nutt, E.C.: Making Tough Decisions: Tactics for Improving Managerial Decision Making. Jossey Bass Pub., San Francisco (1988)
15. Driver, M.J.: The Dynamic Decision Maker. Harper & Row, New York (1990)
16. Scott, S.G., Bruce, R.A.: Decision Making Style: The Development and Assessment of a New Measure, Education and Psychological Measurement (in Press 1995)
17. Schiaffino, S., Amandi, A.: User-interface agent interaction: personalization issues. *Int. J. Hum.-Comput. Stud.* **60**(1), 129–148 (2004)
18. Wang, M.: Design of Weapon Control Interface Based on Human Factors. Beijing Institute of Technology (2015)
19. Zhong, Y., Zhang, C., Shi, Z.: Research on the fuzzy reasoning car - following model considering driver's behavior. Research on the improvement of fuzzy inference following model considering driver behavior. In: *Traffic Information and Security* **28**(3), 17–20 (2010)
20. Wang, J., Zhengding, L.: The method of determining membership function in fuzzy control. *Henan Sci.* **18**(4), 348–351 (2000)
21. Qiongfang, Y., Chen, Y.: Constructing strategy of membership function in fuzzy mathematics. *J. Luohe Vocat. Tech. Coll. (Compr.)* **2**(1), 12–14 (2003)
22. Wang, H., Zhuang, Z.: Determination of membership function in fuzzy reliability analysis. *Electron. Prod. Reliab. Environ. Test.* **8**(4), 2–7 (2000)
23. Li, J., Li, Y.: A further discussion on the determination of membership functions. *J. Guizhou Univ. Technol. Nat. Sci. Ed.* **33**(6), 1–4 (2004)
24. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. *Science* **344** (6191), 1492–1496 (2014)

Driver's Multi-Attribute Task Battery Performance and Attentional Switch Cost Are Correlated with Speeding Behavior in Simulated Driving

Jie Zhang^{1,2}, Mengnuo Dai^{1,2}, and Feng Du^{1,2}(✉)

¹ Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Science, Beijing, People's Republic of China
duf@psych.ac.cn

² Chinese Academy of Sciences, Beijing, People's Republic of China

Abstract. Speeding is one of the leading factors for traffic casualties. It is important to identify underlying factors related with speeding behavior. Present study aimed to explore the relationship between speeding and two general cognitive abilities: multi-tasking and attention-switching abilities. We measured multi-tasking ability using Multi-Attribute Task Battery (MATB). The MATB performance includes hit rate and RT for monitoring task, track error for tracking task and control rate for resource management task. We used the attentional blink (AB) task to measure attention-switching ability. The AB refers to people's inability to detect a second target (T2) that follows within about five hundred milliseconds of an earlier target (T1) in the same location. The attentional switch cost, specifically AB magnitude, is the difference between the highest and lowest accuracy of T2 given correct report of T1 across five T1-T2 intervals. Finally, a driving simulator was used to measure drivers' speeding behavior. The results showed (1) max speeding ratio was significantly correlated with RT for monitoring task, control rate for resource management and AB magnitude; (2) regression analysis show that MATB performance and Attentional switch cost played the key role in predicting max speeding ratio while controlling the demographic variables, but only MATB performance had a significant effect on speeding duration. Thus MATB performance and attentional switch costs is important to predict speeding behavior in simulated driving.

Keywords: Attentional switch cost · Attentional blink · Multi-Attribute Task Battery · Simulated driving · Speeding

1 Introduction

Speeding is one of the leading factors for traffic casualties [1, 2]. In China, official statistics show that in 2013 there were 198394 recorded traffic crashes that resulted in 272263 casualties, of which nearly six percent were caused by speeding [3]. Speeding not only increases crash risks but also affects the severity of a crash [4]. A case-control study conducted by Kloeden et al. showed that the speed-crash rate relationship

followed an exponential function on rural road with speed limits between 80 and 120 km/h [5]. Besides, Miltner and Salvender found fatality risk for belted front-seat passenger was about 30 times higher at 80 km/h than at 40 km/h [6]. Hence, it is important to identify underlying factors related with speeding behaviors.

Elandar et al. proposed a fourfold classification for variables that are related to crash risk: driving skills, driving styles, extrinsic abilities and traits [7]. The extrinsic abilities referred to those general perceptual-motor skills which play key roles in driving safety but extend beyond driving skills. Researchers found that ability to detect visual signals embedded in a complex background and ability to switch attention rapidly are related with better driving safety [8–11]. Present study aimed to examine the relationship between speeding and two general abilities: multi-tasking and attention-switching abilities.

Proper speed control requires drivers to simultaneously monitor car dashboard and road condition. In order to avoid collision and speed violation, drivers have to keep a safe distance from pedestrians, vehicles and any other potentially hazardous obstacles on the road while maintain their speed under limits. Therefore we proposed that multi-tasking ability is critical for speed control. We measured multi-tasking ability using Multi-Attribute Task Battery (MATB) which provides a set of simulated aviation tasks for laboratory studies [12]. The MATB requires operator to continuously track a randomly moving target (tracking task) while monitoring several warning lights and gauges (monitoring task), and managing fuel level in a simulated dynamic fuel system (resource management task). The three tasks simulate what aircrews regularly perform in their real-world task.

Attention-switching ability have been shown to be critical for driving safety because drivers need to constantly switch their attention between road situation and dashboard. In previous studies, the Visual Selective Attention Test (VSAT) was used to test the ability of switching attention spatially. The VSAT involves simultaneous presentation of two streams of numbers and letters at two sides of a screen. Participants are instructed to respond to certain stimuli in the two streams (e.g., all odd numbers at the left and even numbers at the right) according to the cue preceded to the beginning of the streams [13]. However, rapid response on multiple visual items (e.g., road condition, traffic signs, dashboard), which is essential to speed control, relies heavily on not only the spatial attention-switching ability but also the ability to process items rapidly, particularly the temporal attentional-switching ability. In this study, we tested whether temporal attentional-switching ability was a key factor for speed control by using attentional blink (AB) task. The AB refers to people's inability to detect or identify a second target (T2) that follows within about five hundred milliseconds of an earlier target (T1) in the same location [14–16]. Less switch cost in AB reflects better temporal attention-switching ability [17].

Driving simulators provide a controllable, cost-effective and safe testing environment for dangerous driving behavior [18]. Thus, a driving simulator was used to measure drivers' speeding behavior. MATB performance and attentional switch costs was expected to be correlated with speeding behavior in simulated driving task.

2 Method

2.1 Participants

37 participants (22 males and 15 females) took part in this experiment. They ranged in age from 22 to 50 years, with an average age of 29.9 years. All had normal or corrected-to-normal vision, valid driver’s licenses and at least one year of driving experience.

2.2 Apparatus

The MATB and the AB task were implemented on a Core i7 desktop computer equipped with a 17-in. CRT monitor, a joystick and a standard keyboard. The monitor had a refresh rate of 85 Hz, a resolution of 1024 × 768 pixels and a viewing distance of 60 cm. The Sim-Trainer driving simulator, manufactured by Beijing Sunheart Inc., was used for the simulated driving task. The simulator consists of a complete cockpit and three high resolution displays, providing a 120° field of view.

2.3 MATB

We adopted the monitoring, tracking and resource management tasks from MATB to measure multi-tasking ability (Fig. 1).

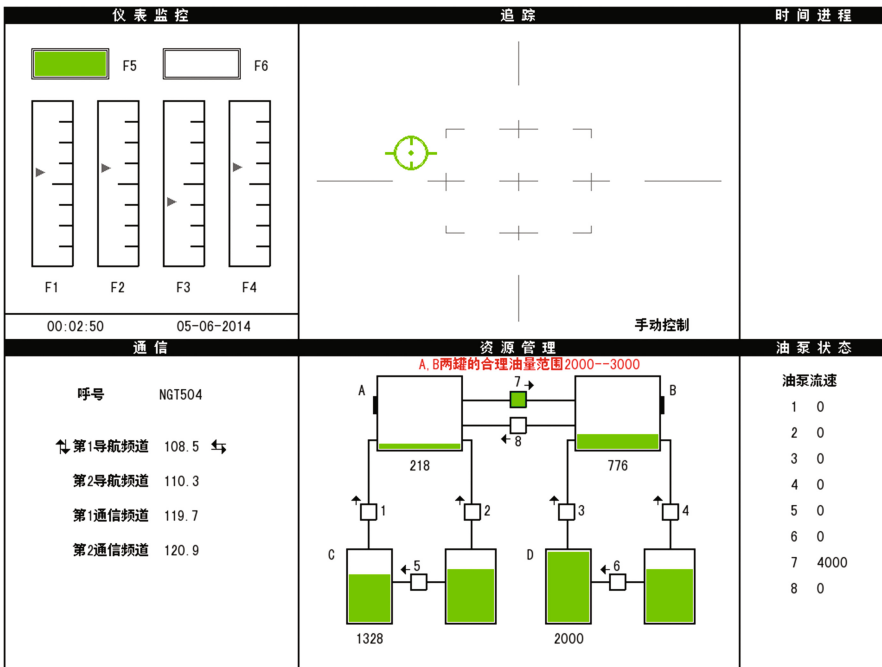


Fig. 1. Illustration of the interface in the Multi-Attribute Task Battery (MATB) (Color figure online)

The monitoring required attending to the four vertical gauges and the two warning lights (the upper left corner of display). In normal condition, the Green light (marked F5) was on, the Red light (marked F6) was off and the pointers of the gauges were within one unit above or below the centers. Participants were instructed to respond as soon as possible to the absence of the Green light, the presence of the Red light and the abnormally large deviation of the pointers by pressing corresponding keys on the joystick. The abnormal status of gauges and warning lights were randomly arranged and were counted as the abnormalities of monitoring of which the total number was 10 and 24 in two experimental blocks. The participant's hit rate and reaction time were calculated and recorded.

As the upper right corner of display shows, participant needed to keep tracking of a randomly-moving target in the tracking task. The root mean square (RMS) track errors, which was deviation from center of tracking target in pixel units, was recorded every 2.4 s. The tracking task were identical in two experimental blocks.

The resource management task required operator to maintain both tank A and B within the range of 2000–3000 units, which was indicated graphically by two black bars on the sides of the two tanks. This was done by turning on or off any of the eight pumps through pressing the corresponding keys on the joystick. All pumps were off at the onset of the task. Both tank A and B had 2100 units of fuel at the beginning and were depleted of fuel at the rate of 800 units per minute. The status of tank A and B were recorded every 14 ms. The parameters of the resource management task were identical in two experimental blocks. The control rate of tank A and tank B, the time percentage when target tank was in the desired range, were calculated.

2.4 Attentional Blink

In AB task, Participants were required to report the two targets embedded in a RSVP stream. After reporting the first target (T1) correctly, participants usually have difficulty in identifying the second target (T2). The impairment for reporting T2 is attentional blink. The attentional switch cost is the AB magnitude which is the difference between the highest and lowest accuracy of T2 given correct report of T1 across five T1-T2 intervals for each participant.

The RSVP stream was presented at the center of display. The background of the screen was black. Each trial began with the presentation of a fixation cross at the center of the screen. After 600 ms, the fixation cross was replaced by a rapidly-changing letter stream consisting of 20 upper-case white letters (1.3° in height). Letters were randomly chosen from the alphabet except the letter I. Each of the letters was presented for 40 ms and was followed by a 40 ms black screen interval, making the SOA 80 ms. T1, the first target, was a white digit randomly chosen between 2 and 9. It could appear in the 10th, 11th, or 12th frame in the stream. The letters kept changing at the same rate after T1 was presented. T2, the second target appeared in the 1st, 2nd, 3rd, 4th or 5th frames after T1. The T2 was a white letter chosen from letter A, B, X or Y. Participants were instructed to report both T1 and T2 as accurately as possible.

2.5 Simulated-Driving Task

We measured speeding behavior based on a simulated-driving task originated from our previous study [19]. Guided by auditory instructions, participants drove along a 3.6 km urban road on the driving simulator. Participants were instructed to limit their speed according to the speed signs. If they exceeded, the simulator would record the speeding duration and calculate the max speeding ratio as the max ratio of speed to speed limit.

2.6 Procedure

Participants came to lab twice at the interval of one week to avoid fatigue effect. Half participants first completed the AB and the MATB, and the other first completed the AB and the simulated-driving task.

The MATB began with 4 practice blocks: each block lasted 5 min, the first three blocks contained only one of the three sub-tasks without repetition and the last block contained both resource management and tracking task. Before the experimental blocks, participants were instructed that they would be performing the monitoring task, the tracking task and the resource management task simultaneously. There were 2 experimental blocks, each session lasted 5 min. The number of abnormalities in the monitoring task was randomly assigned to the two experimental blocks. The AB task consisted of 16 practice trails and 160 experimental trails. The simulated-driving task began with a 5 min practice session in which participants drove freely in a city to get accustomed the simulator. Before the experimental session, participants were instructed to limit their speed according to the speed signs and follow auditory instructions. The experimental session lasted 12–15 min.

3 Results

3.1 Task Performance

There was no practice effect in the MATB, therefore the data of the two experimental blocks of the MATB were combined. Furthermore, the control rate of the two target tanks in the resource management were merged because no significant difference between the target tanks was observed.

Table 1 listed the four MATB indices from the three sub-tasks, the AB magnitude, and two speeding indices from the simulated driving task. An additional correlation analysis between the MATB indices revealed strong positive correlations between the performance of the three sub-tasks: (1) the correlation between RT for monitoring task and track error for tracking task was significant ($r = .47$, $p < .01$), with slower anomaly detection corresponding to larger track deviation; (2) control rate for resource management was significantly correlated with tracking error, ($r = -.60$, $p < .01$), with poorer fuel management corresponding to larger track deviation; (3) control rate for resource management had a significant correlation with RT for monitoring task ($r = -.70$, $p < .01$), with poorer fuel management corresponding to slower anomaly detection.

Table 1. The MATB performance, AB magnitude and speeding indices (Mean ± SE)

MATB				AB	Simulated driving	
HR	RT (s)	TE (pixel)	CR	M	MSR	SD(s)
0.93 ± 0.01	3.69 ± 0.20	78.64 ± 3.74	0.87 ± 0.01	0.30 ± 0.02	0.29 ± 0.03	72.14 ± 6.61

Not. HR – Hit rate for monitoring; RT – RT for monitoring; TE – Track error for tracking; CR – Control rate for resource management; M – magnitude of AB; MSR – Max speeding ratio; SD – Speeding duration.

As shown in Fig. 2, the expected attentional blink was observed. The average T2 accuracy given correct reaction of T1 (T2/T1) reached its lowest at lag3, which was significantly lower than the average T2/T1 at lag1 ($t(36) = 2.31, p < .05$). Moreover, the AB magnitude, shown in Table 1, was calculated as the difference between the highest and lowest accuracy of T2/T1 across five T1-T2 intervals for each participant.

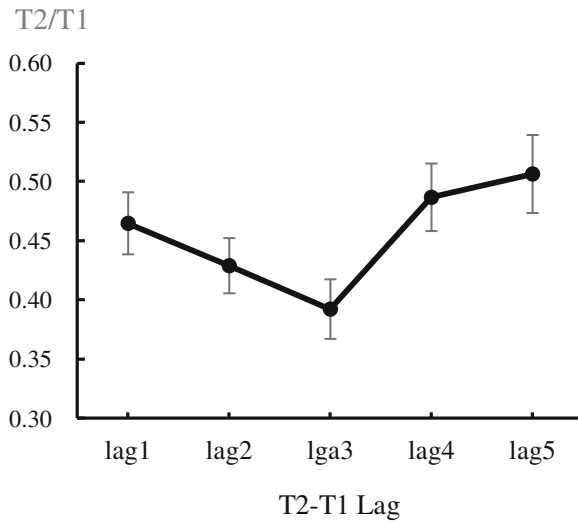


Fig. 2. The average T2/T1 as a function of T2-T1 lags

3.2 Correlation Analysis

The correlations between MATB performance, AB magnitude and speeding indices of the simulated driving task are shown in Table 2. The results showed, (1) monitoring RT was significantly correlated with max speeding ratio ($r = .32, p < .05$), with slower anomaly detection corresponding to higher max speeding ratio; (2) control rate of resource management had a significant negative correlation with max speeding ratio ($r = -.40, p < .05$), with worse resource management rate corresponding to higher max speeding ratio; (3) AB magnitude was significantly correlated with max speeding ratio ($r = .46, p < .01$), with larger attentional switch cost corresponding to higher max speeding ratio. No significant correlation for speeding duration was found.

Table 2. The correlation between MATB performance, AB magnitude and speeding indices

	MATB				AB
	HR	RT	TE	CR	M
Max speeding ratio	-.25	.32*	-.15	-.40*	.46**
Speeding duration	-.20	.27	-.30	-.23	.21

Not. HR – Hit rate for monitoring; RT – RT for monitoring; TE – Track error for tracking; CR – Control rate for resource management; M – magnitude of AB; *p < .05, **p < .01.

3.3 Regression Analysis

We separately conducted three-step hierarchical regression analyses on the max speeding ratio and the speeding duration. Age and gender were entered at Step 1 to control potential demographic effect in the prediction of speeding behavior. The MATB performance was entered at Step 2 and finally the AB magnitude was entered at Step 3.

The results of the hierarchical regressions are shown in Table 3. The MATB performance accounted for 19% of the variance in max speeding ratio that is over and above the variance accounted for by age and gender, and this finding was a statistically significant increase. Moreover, in Step 3 the AB magnitude significantly increased R² by 14%. In Step 3 of predicting max speeding ratio, the standardized regression coefficients were significant for tracking error ($\beta = -.50, p < .01$), resource control rate ($\beta = -.54, p < .05$) and AB magnitude ($\beta = .39, p < .01$). The results indicated that drivers with smaller track deviation, poorer resource control rate and larger attentional switch cost have higher max speeding ratio; however the influence that the attentional switch cost has on max speeding ratio is relatively smaller.

Table 3. Predicting max speeding ratio and speeding duration

Step	Overall model		Predictors (β)						
	ΔR^2	ΔF	Age	Gender	HR	RT	TE	CR	M
Max speeding ratio									
1	.18	4.82*	-.24	-.45**					
2	.19	3.48*	-.12	-.25	-.05	-.03	-.52**	-.60*	
3	.14	9.71**	-.02	-.26	.01	.04	-.50**	-.54*	.39**
Speeding duration									
1	.02	1.30	-.13	-.26					
2	.26	4.00**	.03	-.03	-.18	.18	-.72**	-.39	
3	.00	1.14	.07	-.03	-.15	.19	-.70**	-.37	.16

Not. HR – Hit rate for monitoring; RT – RT for monitoring; TE – Track error for tracking; CR – Control rate for resource management; M – magnitude of AB; *p < .05, **p < .01.

In the regression model of speeding duration, the MATB performance accounted for 26% of the variance in speeding duration that is a significant increase over and above the variance accounted for by age and gender, while the entrance of the AB magnitude made no significant difference in the model. The standardized regression coefficient was significant only for track error ($\beta = -.70$, $p < .01$). The result indicated that when other variables are controlled, smaller track deviation predicts longer speeding duration.

4 Discussion and Conclusion

The purpose of this study was to examine the relationship between speeding behavior and two cognitive abilities including multi-tasking and attention-switching abilities. The results show that the MATB performance and AB magnitude can predict the speeding behavior in a simulated driving.

The present study found a significant positive correlation between the monitoring RT in MATB and the max speeding ratio of simulated driving. This finding suggests that the monitoring task overlaps with the speed control task. The monitoring task requires continuously monitoring the warning lights and abnormal situation of gauges scales [12], whereas speed control requires continuously monitoring the driving speed of a car and its trajectory. Moreover, the resource control rate of resource management task was significantly correlated with the max speeding ratio in simulated driving task. This finding indicates that the resource management task also shares common features with the speed control task. Both tasks require sustaining a high level of vigilance and choosing appropriate pumps/pedal strategically [12]. Finally, the correlation between the AB magnitude and the max speeding ratio was significant, which suggests that the temporal attention-switching ability plays a key role in speed control.

The hierarchical regression analyses showed that both MATB performance and attentional switch cost were important predictors for max speeding ratio, but only MATB performance had significant effect on speeding duration. In accordance with the correlation analysis, the hierarchical regression found the control rate for resource management and the AB magnitude were predictive of the max speeding ratio when demographic factors were controlled. This finding supports the notion that the control rate for resource management and the AB magnitude are critical factors in predicting speeding behavior in simulated driving. It is worth noting that the standard regression coefficients of tracking error were significantly negative in the regression models for both max speeding ratio and speeding duration, with better tracking performance associated with more speeding behavior. One possible explanation for current results is that drivers who had high tracking performance in multi-tasking might be more confident about their driving skill hence tended to maintain higher speed than what was required. One study showed that self-perception of skill and confidence were strong predictors for speeding behavior [20]. We propose that the relationship between tracking performance and speeding behavior need further investigation and should be understood in the context of multi-tasking.

One limitation of this study is the lack of validation of our measures of speeding behavior. As Godley et al. pointed out, participants generally drove faster in the

instrumented car than the simulator, resulting in absolute validity not being established [21]. Further studies are needed to identify key cognitive factors for actual speeding behavior.

Taken together, this study suggests that MATB performance and attentional switch costs might be useful for predicting speeding behavior in simulated driving.

Acknowledgements. This study was supported by grants from the National Natural Science Foundation of China (31470982), and the scientific foundation of the Institute of Psychology, Chinese Academy of Sciences (Y4CX033008).

References

1. Liselotte, L.A., Glad, L.A., Beilinson, L.: Observed vehicle speed and drivers' perceived speed of others. *Appl. Psychol.* **46**, 287–302 (1997). doi:[10.1080/026999497378377](https://doi.org/10.1080/026999497378377)
2. Elvik, R.: Speed limits, enforcement, and health consequences. *Annu. Rev. Publ. Health* **33**, 225–238 (2012). doi:[10.1146/annurev-publhealth-031811-124634](https://doi.org/10.1146/annurev-publhealth-031811-124634)
3. China Road Traffic Accident Statistics Bureau (CRTASB): China Road Traffic Accidents Statistics. Traffic Administration Bureau of China State Security Ministry, Beijing, China (2013). (in Chinese)
4. Aarts, L., Schagen, I.V.: Driving speed and the risk of road crashes: a review. *Accid. Anal. Prev.* **38**, 215–224 (2006). doi:[10.1016/j.aap.2005.07.004](https://doi.org/10.1016/j.aap.2005.07.004)
5. Kloeden, C.N., Ponte, G., McLean, A.J.: Travelling speed and the risk of crash involvement on rural roads. Australian Transport Safety Board, Canberra, A.C.T. (2001)
6. Miltner, E., Salwender, H.-J.: Influencing factors on the injury severity of restrained front seat occupants in car-to-car head-on collisions. *Accid. Anal. Prev.* **27**, 143–150 (1995). doi:[10.1016/0001-4575\(94\)00039-o](https://doi.org/10.1016/0001-4575(94)00039-o)
7. Elander, J., West, R., French, D.: Behavioral correlates of individual differences in road-traffic crash risk: an examination of methods and findings. *Psychol. Bull.* **113**, 279–294 (1993). doi:[10.1037//0033-2909.113.2.279](https://doi.org/10.1037//0033-2909.113.2.279)
8. Loo, R.: Individual differences and the perception of traffic signs. *Hum. Factors* **20**, 65–74 (1978). doi:[10.1177/001872087802000109](https://doi.org/10.1177/001872087802000109)
9. Barrett, G.V., Thornton, C.L.: Relationship between perceptual style and driver reaction to an emergency situation. *J. Appl. Psychol.* **52**, 169–176 (1968). doi:[10.1037/h0025658](https://doi.org/10.1037/h0025658)
10. Mihal, W.L., Barrett, G.V.: Individual differences in perceptual information processing and their relation to automobile accident involvement. *J. Appl. Psychol.* **61**, 229–233 (1976). doi:[10.1037//0021-9010.61.2.229](https://doi.org/10.1037//0021-9010.61.2.229)
11. Kahneman, D., Ben-Ishai, R., Lotan, M.: Relation of a test of attention to road accidents. *J. Appl. Psychol.* **58**, 113–115 (1973). doi:[10.1037/h0035426](https://doi.org/10.1037/h0035426)
12. Comstock, J.R., Arnegard, R.J.: The multi-attribute task battery for human operator workload and strategic behavior research. National Aeronautics and Space Administration, Langley Research Center, Hampton, VA (1992)
13. Avolio, B.J., Alexander, R.A., Barrett, G.V., Sterns, H.L.: Designing a measure of visual selective attention to assess individual differences in information processing. *Appl. Psychol. Meas.* **5**, 29–42 (1981). doi:[10.1177/014662168100500105](https://doi.org/10.1177/014662168100500105)
14. Chun, M.M., Potter, M.C.: A two-stage model for multiple target detection in rapid serial visual presentation. *J. Exp. Psychol. Hum.* **21**, 109–127 (1995). doi:[10.1037//0096-1523.21.1.109](https://doi.org/10.1037//0096-1523.21.1.109)

15. Raymond, J.E., Shapiro, K.L., Arnell, K.M.: Similarity determines the attentional blink. *J. Exp. Psychol. Hum.* **21**, 653–662 (1995). doi:[10.1037//0096-1523.21.3.653](https://doi.org/10.1037//0096-1523.21.3.653)
16. Shapiro, K.L., Caldwell, J., Sorensen, R.E.: Personal names and the attentional blink: a visual “cocktail party” effect. *J. Exp. Psychol. Hum.* **23**, 504–514 (1997). doi:[10.1037//0096-1523.23.2.504](https://doi.org/10.1037//0096-1523.23.2.504)
17. Green, C.S., Bavelier, D.: Action video game modifies visual selective attention. *Nature* **423**, 534–537 (2003). doi:[10.1038/nature01647](https://doi.org/10.1038/nature01647)
18. Ivancic, K., Hesketh, B.: Learning from errors in a driving simulation: effects on driving skill and self-confidence. *Ergonomics* **43**, 1966–1984 (2000). doi:[10.1080/00140130050201427](https://doi.org/10.1080/00140130050201427)
19. Yuan, Y., Du, F., Qu, W., Zhao, W., Zhang, K.: Identifying risky drivers with simulated driving. *Traffic Inj. Prev.* **17**, 44–50 (2015). doi:[10.1080/15389588.2015.1033056](https://doi.org/10.1080/15389588.2015.1033056)
20. Adams-Guppy, J.R., Guppy, A.: Speeding in relation to perceptions of risk, utility and driving style by British company car drivers. *Ergonomics* **38**, 2525–2535 (1995). doi:[10.1080/00140139508925284](https://doi.org/10.1080/00140139508925284)
21. Godley, S.T., Triggs, T.J., Fildes, B.N.: Driving simulator validation for speed research. *Accid. Anal. Prev.* **34**, 589–600 (2002). doi:[10.1016/s0001-4575\(01\)00056-2](https://doi.org/10.1016/s0001-4575(01)00056-2)

Author Index

- Ahn, Jinho II-57
Anderson, Nathan I-266
Ando, Satoshi I-209
Arend, Matthias G. II-363
Arrabito, Robert I-329
Ashcraft, Chace I-266
Attig, Christiane II-3
Ayaz, Hasan I-106
- Babiloni, Fabio I-87
Bach, Cedric I-247
Banbury, Simon I-329
Barbosa, Débora Nice Ferrari I-164
Behymer, Kyle J. II-15
Berberian, Bruno I-87
Biondi, Francesco II-329
Braithwaite, Graham I-21, I-301
Brüggemann, Bernd II-159
- Cai, Jiayi II-191
Calhoun, Gloria L. II-15
Calvet, Guillaume I-247
Cao, Jiaqi I-21
Cardoso, Caroline de Oliveira I-164
Charron, Mario I-329
Chen, Dong II-191
Chen, Mo II-211
Chen, Qiugu II-133
Chen, Yingchun I-232
Chen, Zi-li II-284
Cooper, Joel II-329
Crandall, Jacob W. I-266
Cummings, M.L. II-393
Curtin, Adrian I-106
- Dai, Mengnuo II-426
de Borba Campos, Márcia II-115
De Crescenzo, Francesca I-87
Di Flumeri, Gianluca I-87
Dong, Chuanting II-297
Dong, Dayong I-232
Dong, Wenjun I-232
Dou, Jinhua II-35
- Du, Feng II-426
Dunbar, Jerone II-339
- Feng, Chuanyan II-306
Forch, Valentin II-45
Franke, Thomas II-3, II-45, II-363
Friedrich, Maik I-285
Fu, Shan I-32, I-64
Funada, Mariko I-117
Funada, Tadashi I-117
- Gao, Qin I-310, I-319
Gauci, Jason II-200
Gilbert, Juan E. II-339
Goethe, Rachel II-329
Goodrich, Michael A. I-266
Greaves, Matthew I-21, I-301
Grosh, John I-266
Guo, Qi II-86, II-211
- Han, Xuan II-413
Hao, Jia II-413
Harada, Etsuko T. I-209
Harris, Don II-222
Hasselberg, Andreas I-87
Heffner, Kevin I-329
Henderson, Jonah I-266
Honecker, Fabian II-231
Hou, Ming I-329
Huang, Dan I-32, I-64
Huang, Shoupeng II-306
Huang, Weifen II-191
Huang-fu, Guang-xia II-284
- Igarashi, Yoshihide I-117
- Jakobi, Jörn I-285
Janneck, Monique I-180
Jiang, Xiang I-319
Jin, Xiaoping I-3
Jing, Haipeng II-191
- Kabuss, Wolfgang I-128
Kearney, Peter I-301

- Kim, Bomyeong II-57
 Kim, Kyungdoh II-57
 Komatsubara, Akinori II-101
 Krems, Josef F. II-3, II-45
 Kwee-Meier, Sonja Th. I-128, II-71
- Lassen, Christian II-159
 Li, An II-179
 Li, Haifeng I-155
 Li, Haoyang I-3
 Li, Hongxia I-141
 Li, Jie II-191
 Li, Jingwen I-343
 Li, Tao I-76, II-317
 Li, Wen-Chin I-21, I-301
 Li, Xiaomei I-155
 Li, Zhizhong I-42, I-406
 Lian, Haitao I-319
 Lin, Jr-Hung I-21
 Lin, Yun II-86, II-211
 Liu, Shuang II-133
 Liu, Wang II-191
 Liu, Wenmeng I-32
 Liu, Zhen II-284
 Lu, Yanyu I-32
- Ma, Guang-fu II-284
 Ma, Yahui I-319
 Ma, Yiran I-310
 Maeda, Yoshitaka II-101
 McCarthy, Pete II-251
 McClellan, Joshua I-266
 McColl, Derek I-329
 Mertens, Alexander I-128, II-71
 Miyashiro, Kozue I-209
 Mossmann, João Batista I-164
 Müller, Nicholas H. I-220
 Muscat, Alan II-200
- Neupane, Aadesh I-266
 Niels, Adelka I-180
 Niu, Yafeng II-86, II-179, II-211
 Noah, Benjamin I-343
- Ohmori, Maito I-209
 Ohneiser, Oliver I-87
 Oppenheim, Ilit I-362
 Oron-Gilad, Tal I-362
- Pan, Dan I-42
 Papenfuß, Anne I-285
 Paz Fonseca, Rochele I-164
 Pei, Yeqing I-3
 Peng, Ningyue II-86
 Perret, Viviane I-247
 Pichelmann, Stefan I-285
 Powale, Pallavi II-393
 Proctor, Robert W. II-57
- Qi, Yannan I-382
 Qin, Jingyan II-35
 Qu, Xingda I-76, II-133
- Rauh, Nadine II-3, II-45
 Reategui, Eliseo Berni I-164
 Remmersmann, Thomas II-159
 Ren, Yong II-297
 Roth, Gunar II-375
 Rothrock, Ling I-343
 Rothwell, Clayton D. II-15
 Rudnick, Georg I-394
 Ruf, Christian I-51
 Ruff, Heath A. II-15
- Schlick, Christopher M. I-128
 Schmitt, Fabian II-375
 Schmitt, Marc II-266
 Schuh, Ånderson II-115
 Schulte, Axel I-394, II-231, II-375
 She, Manrong I-406
 Solovey, Erin T. II-393
 Song, Fei I-310
 Stanton, Neville A. II-363
 Strayer, David II-329
 Stütz, Peter I-51, II-266
 Sun, Hui II-297
 Sun, Rui-shan II-284
 Sun, Xiaofang I-200
 Suzuki, Satoshi II-101
- Tao, Da II-133
 Taylor, Glenn II-145
 Teo, Guan Kiat II-251
 Tian, Sen I-64
 Tian, Yu II-306
 Tomita, Akitoshi I-209
 Trautwein, Irmtrud II-159
 Truschzinski, Martina I-220

- Valadares, Vitor Caetano Silveira I-164
Valtin, Georg I-220
- Wang, Lei II-297
Wang, Lijing I-232
Wang, Lin I-64
Wang, Qiuyu I-319
Wang, Xinglong I-382
Wang, Yanlei II-191
Wang, Yanlong I-232
Wang, Yufan I-310
Wanyan, Xiaoru II-306
Westhoven, Martin II-159
Wiessmann, Marion II-71
Wu, Di II-284
Wu, Jianhui I-200
Wu, Xu II-306
- Xue, Chengqi II-86, II-179, II-211
Xue, Hongjun I-76, II-317
- Xue, Qing II-413
Xuereb, Matthew II-200
- Yano, Hiroaki I-209
Yao, Zhuxi I-200
Yuan, Juan II-133
Yuan, Yi I-200
- Zammit-mangion, David II-200
Zhang, Jie II-426
Zhang, Jing II-86, II-179
Zhang, Kan I-200
Zhang, Qian II-413
Zhang, Xiang II-191
Zhang, Xiaoyan I-76, II-133, II-317
Zhang, Xingjian I-382
Zhang, Yijing I-42
Zheng, Bowen I-3
Zhou, Nan I-141
Zhou, Xiaozhou II-179