LNAI 6781

Don Harris (Ed.)

Engineering Psychology and Cognitive Ergonomics

9th International Conference, EPCE 2011
Held as Part of HCI International 2011
Orlando, FL, USA, July 2011, Proceedings

HCI2011
INTERNATIONAL

Springer

Lecture Notes in Artificial Intelligence 6781
Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Don Harris (Ed.)

Engineering Psychology and Cognitive Ergonomics

9th International Conference, EPCE 2011
Held as Part of HCI International 2011
Orlando, FL, USA, July 9-14, 2011
Proceedings

Springer

Volume Editor

Don Harris
HFI Solutions Ltd.
Bradgate Road, Bedford MK40 3DE, UK
E-mail: don.harris@hfisolutions.co.uk

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-21740-1 e-ISBN 978-3-642-21741-8
DOI 10.1007/978-3-642-21741-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011929044

CR Subject Classification (1998): I.2.0, I.2, H.5, H.1.2, H.3, H.4.2, I.6, J.2-3

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The 14th International Conference on Human–Computer Interaction, HCI International 2011, was held in Orlando, Florida, USA, July 9–14, 2011, jointly with the Symposium on Human Interface (Japan) 2011, the 9th International Conference on Engineering Psychology and Cognitive Ergonomics, the 6th International Conference on Universal Access in Human–Computer Interaction, the 4th International Conference on Virtual and Mixed Reality, the 4th International Conference on Internationalization, Design and Global Development, the 4th International Conference on Online Communities and Social Computing, the 6th International Conference on Augmented Cognition, the Third International Conference on Digital Human Modeling, the Second International Conference on Human-Centered Design, and the First International Conference on Design, User Experience, and Usability.

A total of 4,039 individuals from academia, research institutes, industry and governmental agencies from 67 countries submitted contributions, and 1,318 papers that were judged to be of high scientific quality were included in the program. These papers address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The papers accepted for presentation 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.

This volume, edited by Don Harris, contains papers in the thematic area of engineering psychology and cognitive ergonomics (EPCE), addressing the following major topics:

- Cognitive and psychological aspects of interaction
- Cognitive aspects of driving
- Cognition and the web
- Cognition and automation
- Security and safety
- Aerospace and military applications

The remaining volumes of the HCI International 2011 Proceedings are:

- Volume 1, LNCS 6761, Human–Computer Interaction—Design and Development Approaches (Part I), edited by Julie A. Jacko
- Volume 2, LNCS 6762, Human–Computer Interaction—Interaction Techniques and Environments (Part II), edited by Julie A. Jacko
- Volume 3, LNCS 6763, Human–Computer Interaction—Towards Mobile and Intelligent Interaction Environments (Part III), edited by Julie A. Jacko
- Volume 4, LNCS 6764, Human–Computer Interaction—Users and Applications (Part IV), edited by Julie A. Jacko
- Volume 5, LNCS 6765, Universal Access in Human–Computer Interaction—Design for All and eInclusion (Part I), edited by Constantine Stephanidis

- Volume 6, LNCS 6766, Universal Access in Human–Computer Interaction—Users Diversity (Part II), edited by Constantine Stephanidis
- Volume 7, LNCS 6767, Universal Access in Human–Computer Interaction—Context Diversity (Part III), edited by Constantine Stephanidis
- Volume 8, LNCS 6768, Universal Access in Human–Computer Interaction—Applications and Services (Part IV), edited by Constantine Stephanidis
- Volume 9, LNCS 6769, Design, User Experience, and Usability—Theory, Methods, Tools and Practice (Part I), edited by Aaron Marcus
- Volume 10, LNCS 6770, Design, User Experience, and Usability—Understanding the User Experience (Part II), edited by Aaron Marcus
- Volume 11, LNCS 6771, Human Interface and the Management of Information—Design and Interaction (Part I), edited by Michael J. Smith and Gavriel Salvendy
- Volume 12, LNCS 6772, Human Interface and the Management of Information—Interacting with Information (Part II), edited by Gavriel Salvendy and Michael J. Smith
- Volume 13, LNCS 6773, Virtual and Mixed Reality—New Trends (Part I), edited by Randall Shumaker
- Volume 14, LNCS 6774, Virtual and Mixed Reality—Systems and Applications (Part II), edited by Randall Shumaker
- Volume 15, LNCS 6775, Internationalization, Design and Global Development, edited by P.L. Patrick Rau
- Volume 16, LNCS 6776, Human-Centered Design, edited by Masaaki Kurosu
- Volume 17, LNCS 6777, Digital Human Modeling, edited by Vincent G. Duffy
- Volume 18, LNCS 6778, Online Communities and Social Computing, edited by A. Ant Ozok and Panayiotis Zaphiris
- Volume 19, LNCS 6779, Ergonomics and Health Aspects of Work with Computers, edited by Michelle M. Robertson
- Volume 20, LNAI 6780, Foundations of Augmented Cognition: Directing the Future of Adaptive Systems, edited by Dylan D. Schmorrow and Cali M. Fidopiastis
- Volume 22, CCIS 173, HCI International 2011 Posters Proceedings (Part I), edited by Constantine Stephanidis
- Volume 23, CCIS 174, HCI International 2011 Posters Proceedings (Part II), edited by Constantine Stephanidis

I would like to thank the Program Chairs and the members of the Program Boards of all Thematic Areas, listed herein, for their contribution to the highest scientific quality and the overall success of the HCI International 2011 Conference.

In addition to the members of the Program Boards, I also wish to thank the following volunteer external reviewers: Roman Vilimek from Germany, Ramalingam Ponnusamy from India, Si Jung "Jun" Kim from the USA, and Ilia Adami, Iosif Klironomos, Vassilis Kouroumalis, George Margetis, and Stavroula Ntoa from Greece.

This conference would not have been possible without the continuous support and advice of the Conference Scientific Advisor, Gavriel Salvendy, as well as the dedicated work and outstanding efforts of the Communications and Exhibition Chair and Editor of HCI International News, Abbas Moallem.

I would also like to thank for their contribution toward the organization of the HCI International 2011 Conference the members of the Human–Computer Interaction Laboratory of ICS-FORTH, and in particular Margherita Antona, George Paparoulis, Maria Pitsoulaki, Stavroula Ntoa, Maria Bouhli and George Kapnas.

July 2011 Constantine Stephanidis

Organization

Ergonomics and Health Aspects of Work with Computers

Program Chair: Michelle M. Robertson

Arne Aarås, Norway
Pascale Carayon, USA
Jason Devereux, UK
Wolfgang Friesdorf, Germany
Martin Helander, Singapore
Ed Israelski, USA
Ben-Tzion Karsh, USA
Waldemar Karwowski, USA
Peter Kern, Germany
Danuta Koradecka, Poland
Nancy Larson, USA
Kari Lindström, Finland
Brenda Lobb, New Zealand
Holger Luczak, Germany
William S. Marras, USA
Aura C. Matias, Philippines
Matthias Rötting, Germany
Michelle L. Rogers, USA
Dominique L. Scapin, France
Lawrence M. Schleifer, USA
Michael J. Smith, USA
Naomi Swanson, USA
Peter Vink, The Netherlands
John Wilson, UK

Human Interface and the Management of Information

Program Chair: Michael J. Smith

Hans-Jörg Bullinger, Germany
Alan Chan, Hong Kong
Shin'ichi Fukuzumi, Japan
Jon R. Gunderson, USA
Michitaka Hirose, Japan
Jhilmil Jain, USA
Yasufumi Kume, Japan
Mark Lehto, USA
Hirohiko Mori, Japan
Fiona Fui-Hoon Nah, USA
Shogo Nishida, Japan
Robert Proctor, USA
Youngho Rhee, Korea
Anxo Cereijo Roibás, UK
Katsunori Shimohara, Japan
Dieter Spath, Germany
Tsutomu Tabe, Japan
Alvaro D. Taveira, USA
Kim-Phuong L. Vu, USA
Tomio Watanabe, Japan
Sakae Yamamoto, Japan
Hidekazu Yoshikawa, Japan
Li Zheng, P. R. China

Human–Computer Interaction

Program Chair: Julie A. Jacko

Sebastiano Bagnara, Italy
Sherry Y. Chen, UK
Marvin J. Dainoff, USA
Jianming Dong, USA
John Eklund, Australia
Xiaowen Fang, USA
Ayse Gurses, USA
Vicki L. Hanson, UK
Sheue-Ling Hwang, Taiwan
Wonil Hwang, Korea
Yong Gu Ji, Korea
Steven A. Landry, USA
Gitte Lindgaard, Canada
Chen Ling, USA
Yan Liu, USA
Chang S. Nam, USA
Celestine A. Ntuen, USA
Philippe Palanque, France
P.L. Patrick Rau, P.R. China
Ling Rothrock, USA
Guangfeng Song, USA
Steffen Staab, Germany
Wan Chul Yoon, Korea
Wenli Zhu, P.R. China

Engineering Psychology and Cognitive Ergonomics

Program Chair: Don Harris

Guy A. Boy, USA
Pietro Carlo Cacciabue, Italy
John Huddlestone, UK
Kenji Itoh, Japan
Hung-Sying Jing, Taiwan
Wen-Chin Li, Taiwan
James T. Luxhøj, USA
Nicolas Marmaras, Greece
Sundaram Narayanan, USA
Mark A. Neerincx, The Netherlands
Jan M. Noyes, UK
Kjell Ohlsson, Sweden
Axel Schulte, Germany
Sarah C. Sharples, UK
Neville A. Stanton, UK
Xianghong Sun, P.R. China
Andrew Thatcher, South Africa
Matthew J.W. Thomas, Australia
Mark Young, UK
Rolf Zon, The Netherlands

Universal Access in Human–Computer Interaction

Program Chair: Constantine Stephanidis

Julio Abascal, Spain
Ray Adams, UK
Elisabeth André, Germany
Margherita Antona, Greece
Chieko Asakawa, Japan
Christian Bühler, Germany
Jerzy Charytonowicz, Poland
Pier Luigi Emiliani, Italy
Michael Fairhurst, UK
Dimitris Grammenos, Greece
Andreas Holzinger, Austria
Simeon Keates, Denmark
Georgios Kouroupetroglou, Greece
Sri Kurniawan, USA
Patrick M. Langdon, UK
Seongil Lee, Korea

Zhengjie Liu, P.R. China
Klaus Miesenberger, Austria
Helen Petrie, UK
Michael Pieper, Germany
Anthony Savidis, Greece
Andrew Sears, USA
Christian Stary, Austria
Hirotada Ueda, Japan
Jean Vanderdonckt, Belgium
Gregg C. Vanderheiden, USA
Gerhard Weber, Germany
Harald Weber, Germany
Panayiotis Zaphiris, Cyprus

Virtual and Mixed Reality

Program Chair: Randall Shumaker

Pat Banerjee, USA
Mark Billinghurst, New Zealand
Charles E. Hughes, USA
Simon Julier, UK
David Kaber, USA
Hirokazu Kato, Japan
Robert S. Kennedy, USA
Young J. Kim, Korea
Ben Lawson, USA
Gordon McK Mair, UK
David Pratt, UK
Albert “Skip” Rizzo, USA
Lawrence Rosenblum, USA
Jose San Martin, Spain
Dieter Schmalstieg, Austria
Dylan Schmorrow, USA
Kay Stanney, USA
Janet Weisenford, USA
Mark Wiederhold, USA

Internationalization, Design and Global Development

Program Chair: P.L. Patrick Rau

Michael L. Best, USA
Alan Chan, Hong Kong
Lin-Lin Chen, Taiwan
Andy M. Dearden, UK
Susan M. Dray, USA
Henry Been-Lirn Duh, Singapore
Vanessa Evers, The Netherlands
Paul Fu, USA
Emilie Gould, USA
Sung H. Han, Korea
Veikko Ikonen, Finland
Toshikazu Kato, Japan
Esin Kiris, USA
Apala Lahiri Chavan, India
James R. Lewis, USA
James J.W. Lin, USA
Rungtai Lin, Taiwan
Zhengjie Liu, P.R. China
Aaron Marcus, USA
Allen E. Milewski, USA
Katsuhiko Ogawa, Japan
Oguzhan Ozcan, Turkey
Girish Prabhu, India
Kerstin Röse, Germany
Supriya Singh, Australia
Alvin W. Yeo, Malaysia
Hsiu-Ping Yueh, Taiwan

Online Communities and Social Computing

Program Chairs: A. Ant Ozok, Panayiotis Zaphiris

Chadia N. Abras, USA
Chee Siang Ang, UK
Peter Day, UK
Fiorella De Cindio, Italy
Heidi Feng, USA
Anita Komlodi, USA
Piet A.M. Kommers, The Netherlands
Andrew Laghos, Cyprus
Stefanie Lindstaedt, Austria
Gabriele Meiselwitz, USA
Hideyuki Nakanishi, Japan
Anthony F. Norcio, USA
Ulrike Pfeil, UK
Elaine M. Raybourn, USA
Douglas Schuler, USA
Gilson Schwartz, Brazil
Laura Slaughter, Norway
Sergei Stafeev, Russia
Asimina Vasalou, UK
June Wei, USA
Haibin Zhu, Canada

Augmented Cognition

Program Chairs: Dylan D. Schmorrow, Cali M. Fidopiastis

Monique Beaudoin, USA
Chris Berka, USA
Joseph Cohn, USA
Martha E. Crosby, USA
Julie Drexler, USA
Ivy Estabrooke, USA
Chris Forsythe, USA
Wai Tat Fu, USA
Marc Grootjen, The Netherlands
Jefferson Grubb, USA
Santosh Mathan, USA
Rob Matthews, Australia
Dennis McBride, USA
Eric Muth, USA
Mark A. Neerincx, The Netherlands
Denise Nicholson, USA
Banu Onaral, USA
Kay Stanney, USA
Roy Stripling, USA
Rob Taylor, UK
Karl van Orden, USA

Digital Human Modeling

Program Chair: Vincent G. Duffy

Karim Abdel-Malek, USA
Giuseppe Andreoni, Italy
Thomas J. Armstrong, USA
Norman I. Badler, USA
Fethi Calisir, Turkey
Daniel Carruth, USA
Keith Case, UK
Julie Charland, Canada
Yaobin Chen, USA
Kathryn Cormican, Ireland
Daniel A. DeLaurentis, USA
Yingzi Du, USA
Okan Ersoy, USA
Enda Fallon, Ireland
Yan Fu, P.R. China
Afzal Godil, USA

Ravindra Goonetilleke, Hong Kong
Anand Gramopadhye, USA
Lars Hanson, Sweden
Pheng Ann Heng, Hong Kong
Bo Hoege, Germany
Hongwei Hsiao, USA
Tianzi Jiang, P.R. China
Nan Kong, USA
Steven A. Landry, USA
Kang Li, USA
Zhizhong Li, P.R. China
Tim Marler, USA
Ahmet F. Ozok, Turkey
Srinivas Peeta, USA
Sudhakar Rajulu, USA
Matthias Rötting, Germany
Matthew Reed, USA
Johan Stahre, Sweden
Mao-Jiun Wang, Taiwan
Xuguang Wang, France
Jingzhou (James) Yang, USA
Gulcin Yucel, Turkey
Tingshao Zhu, P.R. China

Human-Centered Design

Program Chair: Masaaki Kurosu

Julio Abascal, Spain
Simone Barbosa, Brazil
Tomas Berns, Sweden
Nigel Bevan, UK
Torkil Clemmensen, Denmark
Susan M. Dray, USA
Vanessa Evers, The Netherlands
Xiaolan Fu, P.R. China
Yasuhiro Horibe, Japan
Jason Huang, P.R. China
Minna Isomursu, Finland
Timo Jokela, Finland
Mitsuhiko Karashima, Japan
Tadashi Kobayashi, Japan
Seongil Lee, Korea
Kee Yong Lim, Singapore
Zhengjie Liu, P.R. China
Loïc Martínez-Normand, Spain
Monique Noirhomme-Fraiture, Belgium
Philippe Palanque, France
Annelise Mark Pejtersen, Denmark
Kerstin Röse, Germany
Dominique L. Scapin, France
Haruhiko Urokohara, Japan
Gerrit C. van der Veer, The Netherlands
Janet Wesson, South Africa
Toshiki Yamaoka, Japan
Kazuhiko Yamazaki, Japan
Silvia Zimmermann, Switzerland

Design, User Experience, and Usability

Program Chair: Aaron Marcus

Ronald Baecker, Canada
Barbara Ballard, USA
Konrad Baumann, Austria
Arne Berger, Germany
Randolph Bias, USA
Jamie Blustein, Canada
Ana Boa-Ventura, USA
Lorenzo Cantoni, Switzerland
Sameer Chavan, Korea
Wei Ding, USA
Maximilian Eibl, Germany
Zelda Harrison, USA

Rüdiger Heimgärtner, Germany
Brigitte Herrmann, Germany
Sabine Kabel-Eckes, USA
Kaleem Khan, Canada
Jonathan Kies, USA
Jon Kolko, USA
Helga Letowt-Vorbek, South Africa
James Lin, USA
Frazer McKimm, Ireland
Michael Renner, Switzerland
Christine Ronnewinkel, Germany
Elizabeth Rosenzweig, USA
Paul Sherman, USA
Ben Shneiderman, USA
Christian Sturm, Germany
Brian Sullivan, USA
Jaakko Villa, Finland
Michele Visciola, Italy
Susan Weinschenk, USA

HCI International 2013

The 15th International Conference on Human–Computer Interaction, HCI International 2013, will be held jointly with the affiliated conferences in the summer of 2013. It will cover a broad spectrum of themes related to human–computer interaction (HCI), 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 about the topics, as well as the venue and dates of the conference, will be announced through the HCI International Conference series website: http://www.hci-international.org/

General Chair
Professor Constantine Stephanidis
University of Crete and ICS-FORTH
Heraklion, Crete, Greece
Email: cs@ics.forth.gr

Table of Contents

Part I: Cognitive and Psychological Aspects of Interaction

Part II: Cognitive Aspects of Driving

Part III: Cognition and the Web

Part IV: Cognition and Automation

Part V: Security and Safety

Part VI: Aerospace and Military Applications

Part I
Cognitive and Psychological Aspects of Interaction

Movement Time for Different Input Devices

L. Paige Bacon and Kim-Phuong L. Vu

California State University, Long Beach
1250 Bellflower Blvd Long Beach, CA 90840, USA
paigebacon86@gmail.com, kvu8@csulb.edu

Abstract. Fitts' law states that movement time can be predicted by knowing the size of a target to which a person is intending to move and the distance to be moved. The current study measured choice-movement time with three input devices commonly used in human-computer interaction tasks: response panel, computer mouse, and touch-screen. We also examined how direction of movement with the different input devices influences performance. Movement time was shorter when responses were made with the response panel than with the mouse and touch-screen. Furthermore, horizontal movement time was faster than vertical movement time, even when the size of the stimuli and distance to be moved were equal. Fitts' law was used to estimate the slope and intercepts of the functions for each input device and dimension to determine whether the devices and dimensions had greater influence on the starting time or the speed of execution.

Keywords: Fitts' law, input device, movement time, display-control compatibility.

1 Introduction

When a person interacts with his or her daily environment, it is likely that s/he will perform aimed movements. Examples of aimed movements include a) using a computer mouse to move a cursor to a menu item in order to select that item, b) moving a hand from the steering wheel of a vehicle to press a button on the control panel to change the radio station, and c) moving a hand from its current position to touch an icon on the screen of an iPhone to launch an application. Fitts' law states that movement time (*MT*) can be predicted by knowing the size the target and the distance from the target [1]. The shorter the distance and the larger the target, the faster the movement time will be.

Fitts' law refers to a mathematical relationship between movement speed and accuracy of movement. In 1954, Paul Fitts began exploring this relationship by having participants tap a metal, hand-held stylus onto two metal plates as many times as possible in a given period of time. During the experiment, Fitts also manipulated the size of the plates (target width, *W*) and the distance between them (amplitude, *A*), which allowed him to test performance across a variety of combinations of target sizes and distances. The participants were instructed to consider accuracy of movement as a higher priority than speed of movement. Performance was measured by the number of taps that could be completed in a given 15-second trial. Fitts and Peterson [2] continued to investigate this relationship, and concluded that movement

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 3–9, 2011.

time (*MT*) was a function of the difficulty of the movement, expressed as a logarithmic ratio of amplitude and width, known as the index of difficulty (*ID*). *Fitts' law* is expressed as $MT = a + b\ (ID)$, where a and b are empirically derived constants. Fitts' law is important to the field of HCI because many tasks involve aimed movements.

Fitts' law is robust, having been shown to hold for movements of the head [3] and feet [4], as well as for movements made underwater [5] and with remote manipulation [6]. Thus, many designers will benefit from using Fitts' law to estimate movement times. In particular, designers can benefit from using Fitts' law to determine how a device or movement direction affects performance. Because the constants in Fitts' law (a and b) represent properties of the device being tested, knowing these values can help designers diagnose where the cost in movement times with different devices originates. The constant, a (the intercept) is indicative of the time needed to start the movement with the device. Constant b, (the slope) is indicative of the inherent movement speed using the device.

Stimulus-response (S-R) compatibility, or how natural a response to a stimulus is, can also influence the slope of the Fitts' law function [7], where a higher degree of compatibility reduces the slope. There is also some evidence from the S-R compatibility literature suggesting that, in many situations, reaction time is faster for horizontal than vertical S-R relations, a phenomenon known as the right-left prevalence effect [8, 9, 10]. However, it is not known whether the advantage for the horizontal dimension would be evident in movement time. Thus, the goal of the current study was to examine whether the direction of movement in a choice reaction task has an effect on movement time with different input devices used in HCI tasks.

Three input devices were examined: response panel, computer mouse, and touch screen. All response devices were mapped compatibly to the stimulus display. The response devices differed with respect to their integration with the display. The touch screen was the most integrated input device because participants touched the target to make a response. The computer mouse was less integrated because participants moved the mouse to produce cursor movement to a target that was not on the same plane. Finally, the response panel was the least integrated because participants pressed a corresponding button on the response panel that was separate from the display.

2 Method

2.1 Participants

Forty-four undergraduate students (24 women, 20 men), ages 18-39 (*M* = 20.14 years), enrolled in an Introductory Psychology course at California State University, Long Beach participated the study. Participants were recruited from the Psychology Subject Pool and received credit toward their Introductory Psychology requirement. Each participant completed a demographic questionnaire after completing the experimental trials.

Participants reported using a computer an average of 20.5 hours per week. Participants were also asked to use a scale of 1 to 7, with 1 being no experience and 7 being very experienced, to rate their experience with different input devices. Participants rated their experience to be greatest for the computer mouse (M = 6.6) and a computer keyboard (M = 6.5), and least for a touch screen (M = 5.0).

2.2 Design

A 2 (distance: near and far) x 2 (dimension: horizontal and vertical) x 3 (device: mouse, touch screen, and response panel) within-subjects factorial design was used. The device variable was counterbalanced across participants prior to their arrival for the experiment.

2.3 Materials

The experimental apparatus consisted of an Asus Eee Top touch screen computer running Microsoft Windows XP Home Edition, a Microsoft IntelliMouse Explore 3.0, which was used during the mouse device condition. An Ergodex® DX1 Input System Panel was used in the response panel device condition. The response buttons were compatibly mapped with the stimuli that were presented on the screen. A custom program written using Microsoft Visual Basic 2008 Express Edition controlled the experiment and collected the data. The touch screen computer was located on a table that was 154 cm wide, 76 cm deep, with a height of 75 cm from the floor. The near edge of the table was 23 cm from the base of the computer.

2.4 Procedure

The experiment was conducted in a single session lasting approximately 45 minutes. Participants were seated at a table and given two copies of the informed consent form. All agreed to participate and signed the informed consent form, after which the experiment began.

For each condition, upon the press and release of the “Start Trial” button by the participant (located in the middle of the screen; see Figure 1), a target randomly appeared in one of four directions (above, below, left or right) and at one of two distances (near or far) from the “Start Trial” button. Participants were instructed to move to the stimulus once they had identified the target. The trial ended when the participant responded at the location of the target.

Movement time was measured from the time that the participant clicked on the “Start Trial” button until they selected the target. Ten practice trials with each device were given prior to beginning data collection. The experimental block consisted of 40 trials per stimulus location at the two distances (320 trials per device). The target consisted of a 1.27 cm by 1.27 cm black, square that appeared on a tan background. Participants were given rest periods between device conditions and were also told that they could rest at any time during data collection as long as there was no target on the screen. At the completion of the experimental trials, participants completed a demographic questionnaire. Finally, they were debriefed and thanked for their participation.

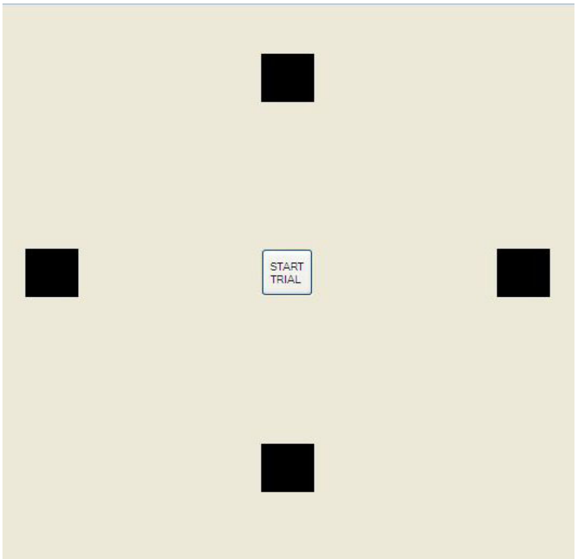


Fig. 1. Task Display. Participants clicked on the start trial button (or a response button located in the same spatial position on a response panel) to start the trial. Then, one target stimulus (depicted by the black squares) appeared, and participants were to respond by moving to the location of the target.

3 Results

A 2 (distance: near and far) x 2 (dimension: horizontal and vertical) x 3 (device: mouse, touch screen, and response panel) analysis of variance was conducted on mean movement time for each participant in each condition.

There was a significant main effect of distance, *F* (1,43) = 398, $p < .001$, $\eta^2 = .90$. Consistent with Fitts' law, movement time was longer for further distances than nearer distances (*M* = 520 ms, *SE* = 15.9; *M* = 400 ms, *SE* = 12.7). A significant main effect of dimension, *F* (1,43) = 20.4, $p < .001$, $\eta^2 = .32$, was also obtained in which movement time for the horizontal dimension was faster (*M* = 450 ms, *SE* = 13.52) than movement time for the vertical dimension (*M* = 472 ms, *SE* = 15.1).

The main of effects of distance and dimension were qualified by a significant distance x dimension interaction, *F* (1,43) = 4.4, $p < .05$, $\eta^2 = .09$, see Figure 2. Tests of simple effects were performed with the Bonferroni correction. Movement time was significantly different across both factors; the shortest movement time occurred when the target was near and in the horizontal direction (*M* = 393 ms, *SE* = 12.4), followed by when the target was near and in the vertical direction (*M* = 409 ms, *SE* = 13.3), followed by when the target was far and in the horizontal direction (*M* = 507 ms, *SE* = 15.4), and was longest when the target was far and in the vertical direction

(*M* = 535.64 ms, *SE* = 17.19), $p < .001$. In other words, this interaction reflects the fact that horizontal movement time was less affected by distance than vertical movement time.

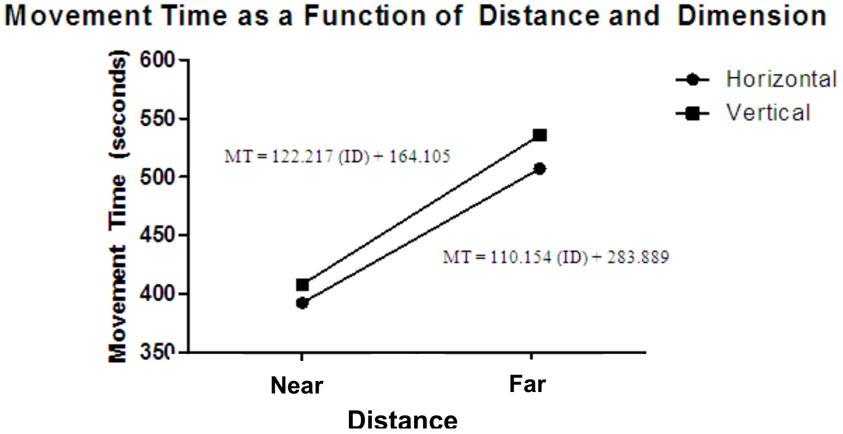


Fig. 2. Movement time as a function of distance and dimension

A significant main effect of input device was also found, *F* (2,42) = 112, $p < .001$, $\eta^2 = .84$, with movement time using the mouse being the longest (*M* = 569 ms, *SE* = 15.5), followed by movement time using the touch screen (*M* = 456 ms, *SE* = 18.4), and movement time using the response panel being the shortest (*M* = 358 ms, *SE* = 14.1). However, this main effect was qualified by a significant distance x device interaction, *F* (2,42) = 39, $p < .001$, $\eta^2 = .66$, see Figure 3. A post-hoc analysis using the Bonferroni correction was conducted to determine if there was a significant difference in the movement time with the different input devices when the target was near compared to when it was far. When the target was near, there was a significant difference in movement time across devices, where movement time was longest using the mouse (*M* = 492 ms, *SE* = 14.7), intermediate when using the touch screen (*M* = 393 ms, *SE* = 15.9), and shortest using the response panel (*M* = 317 ms, *SE* = 13.2). The same pattern held when the distance was far, where movement time for the mouse was still longest (*M* = 647 ms, *SE* = 17.4), followed by the touch screen (*M* = 518 ms, *SE* = 21.3), and response panel (*M* = 399 ms, *SE* = 15.7). However, the difference between devices was larger at the farther distance than at the nearer distance.

Because dimension did not interact with input device, Fitts' law was used to derive the constants *a* and *b* for each dimension (across device) and each input device (across dimension). The equations for estimating MT according to Fitts' law for each dimension and input device are presented in Figures 2 and 3, respectively.

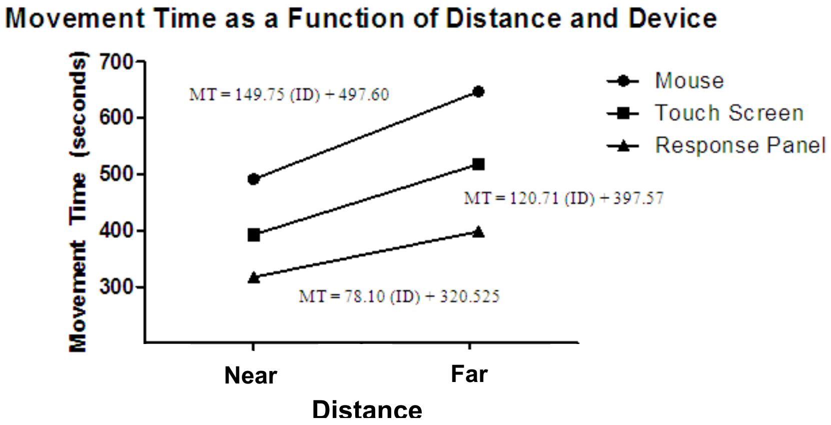


Fig. 3. Movement time as a function of distance and device

4 Discussion

The Fitts' law functions were different for each dimension and input device. For the dimension variable, movement time was found to be shorter for the horizontal than vertical dimension. Upon examining the slope of Fitts' law function, there was an 11% increase from the horizontal dimension to the vertical one. This difference may reflect easier motor control along the horizontal plane than vertical one. However, the increase in the slope of the functions was smaller than the increase observed in the intercept (73%), which is indicative of the time needed to start the movement along the horizontal and vertical dimensions. Because the experimental paradigm was a choice aimed-movement task, the difference may reflect that it takes longer to select responses along the vertical dimension than the horizontal one. This finding is consistent with other choice-reaction studies [11] that showed overall RT for horizontally arrayed S-R sets to be shorter than for vertically arrayed S-R sets.

For the input devices, the response panel yielded the shortest MT, followed by the touch screen, and then the computer mouse. When examining Fitts' law functions, there was a 55% increase in slope of the function when comparing the response panel to the touch screen, but only a 24% increase in the intercept of the functions. Although this finding may suggest that it is easier to move your finger from one key to another than to move your finger from one item on a touch screen to another, it may be the case that occlusion of the display by the hand on the touch screen accounts for part of the slowing. Comparison of the response panel with the mouse resulted in a 92% increase in the slope of the function and a 55% increase in the intercept. Because the two input devices require different types of movement, it is not surprising that most of the increase is in the slope. With the mouse movement, the input device is controlling a cursor on the display. Although the mouse is a zero-order control, in which changes in the position of the mouse correspond to changes in the position of the cursor, the mapping of distance is not direct (e.g., moving the mouse 1 inch to the left may result in the cursor moving 3 inches on the screen). Thus, mouse movement may be longer due to a need to slow the device to hone in on the target or to make corrections required for under- or over-shooting the target.

The implications of the present findings relating to the design of displays and controls are as follows:

- If the environment or situation in which the input will be made requires a short movement time, the design should use a discrete button push for the user to make the input. For example, if a person is driving and would like to change the radio station, it would probably take less time for the movement from the steering wheel to a button on the radio panel than to a button indicator on a touch screen display.
- When designing for a task that requires a short movement time along a single dimension, the design should involve movement along the horizontal dimension instead of the vertical dimension. For example, in a word processing program, such as Microsoft Word, the new ribbon design requiring selection of items from left to right may be more efficient than the old drop-down menu design.

Acknowledgments. This study was supported by NASA cooperative agreement NNX09AU66A, *Group 5 University Research Center: Center for the Human Factors in Advanced Aeronautics Technologies* (Brenda Collins, Technical Monitor).

References

1. Fitts, P.M.: The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. J. Exp. Psychol. 47, 381–391 (1954)
2. Fitts, P.M., Peterson, J.R.: Information Capacity of Discrete Motor Responses. J. Exp. Psychol. 67, 103–112 (1964)
3. Radwin, R.G., Vanderheiden, G.C., Lin, M.-L.: A Method for Evaluating Head-Controlled Input Devices Using Fitts' Law. Hum. Factors 32, 423–438 (1990)
4. Drury, C.G.: Application of Fitts' Law to Foot-Pedal Design. Hum. Factors 17, 368–373 (1985)
5. Kerr, R.: Diving, Adaptation, and Fitts' Law. J. Motor Beh. 10, 255–260 (1978)
6. McGovern, D.E.: Factors Affecting Control Allocation for Augmented Remote Manipulation. Doctoral dissertation, Stanford University (1974)
7. Proctor, R.W., Van Zandt, T.: Human Factors in Simple and Complex Systems. CRC Press, Boca Raton (2008)
8. Proctor, R.W., Vu, K.-P.L.: Stimulus-Response Compatibility: Data, Theory, and Application. CRC Press, Boca Raton (2006)
9. Rubichi, S., Vu, K.-P.L., Nicoletti, R., Proctor, R.W.: Spatial Coding in Two Dimensions. Psych. Bull. & Rev. 13, 201–216 (2006)
10. Nicoletti, R., Umiltà, C.: Right-Left Prevalence in Spatial Compatibility. Percept. & Psychophys. 35, 333–343 (1984)
11. Vu, K.-P.L., Proctor, R.W., Pick, D.F.: Vertical Versus Horizontal Compatibility: Left-Right Prevalence with Bimanual Keypresses. Psychol. Res.-Psych. Fo. 64, 25–40 (2000)

Audio and Audiovisual Cueing in Visual Search: Effects of Target Uncertainty and Auditory Cue Precision

Hugo Bertolotti and Thomas Z. Strybel

California State University, Long Beach,
Center for the Study of Advanced Aeronautics Technologies
1250 N Bellflower Blvd. Long Beach, CA 90840, USA
hbertolo11@yahoo.com, tstrybel@csulb.edu

Abstract. Auditory spatial cue accuracy and target uncertainty were examined within visual search. Participants identified a visual target to be present or absent under various target percentage conditions (25%, 50%, & 100%) with either an auditory cue which was spatially coincident with or displaced 4° or 8° (vertical or horizontal) from the target, or both an auditory and visual cue (circle 6.5° radius; identifying the local-target-area surrounding the target). Within the auditory cue condition, horizontal displacement was a greater detriment to target present search times than vertical displacement, regardless of error magnitude or target percentage. When provided an audiovisual cue, search times decreased 25% for present targets, and as much as 300% for absent targets. Furthermore, within audiovisual cue condition, while present target search times decreased with target percentage, absent target search times increased with target percentage. Cue condition and target uncertainty driven search strategies are discussed, with recommended design requirements and research implementations.

Keywords: Visual Search, Audio Cue, Audiovisual Cue, Auditory Cue Precision, Target Uncertainty, False Alarm.

1 Introduction

In recent decades, technological advancements have made virtual environments and complex audiovisual displays a possible design solution for aircraft cockpits, automobiles, and other complex work environments. Environments such as these require operators to take in and analyze a consistent influx of information and data, and make critical decisions based upon the information. Moreover, with the majority of information presented to the operator being visual, inundation of the already overloaded visual channel is possible. These complex work environments might avoid visual overload by providing some information via other sensory channels, for example, auditory. The information provided from the auditory channel can be either redundant with information presented visually (creating an audiovisual display), or specific to the auditory channel alone (creating an auditory display).

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 10–20, 2011.

Research concerning auditory spatial cueing in visual search has demonstrated several benefits of providing operators with spatially coincident auditory cues –auditory cues provided at a specific elevation and azimuth in space – in both laboratory and applied settings. Within a controlled laboratory environment, coincident auditory spatial cues have been shown to significantly reduce search times required to locate and identify a specified target during a visual search task [1] [2]. In applied settings, simulated 3D auditory spatial cues (simulated interaural difference cues and spectral shape cues provided via headphones) improved target acquisition, traffic detection and avoidance, and visual workload [3] [4] [5]. It is important to note that while providing an auditory spatial cue may be advantageous, the magnitude of the benefits provided is contingent upon the characteristics of the auditory cue and the visual task itself.

One important auditory cue characteristic, auditory cue precision (measured as the distance between the location of the audio cue and the location of the visual target), has been demonstrated to significantly affect target identification [6] [7] [8]. Research has shown that increasing error magnitude displacement (0° to 8°) between a specified target and an auditory spatial cue will significantly increase target search times [7] [8]; however, while search times consistently increased with error magnitude displacement, manipulating error displacement direction produced inconclusive results. Within Vu et al. [6], vertical displacement of an auditory spatial cue was a greater detriment to target identification than horizontal displacement, whereas Bertolotti and Strybel [7] found horizontal displacement to be more detrimental. Moreover, as the error displacement of auditory spatial cue relative to the target increased, the directional discrepancies between Bertolotti and Strybel, and Vu et al. become larger.

A possibility for the inconsistent dimensional findings may be due to target uncertainty. While a target was present on every trial within Vu et al. [6], Bertolotti and Strybel [7] not only included trials which contained no targets (false targets), requiring participants to respond “no target” when none was found, but also manipulated the percentage of trials containing no targets within a given block of trials. When false targets were introduced, horizontal displacement became significantly more detrimental to target identification than vertical displacement, regardless of target percentage. In addition, the time required to identify the absence of a target was three times that of a present target, regardless of cue precision. Perhaps, the introduction of false alarms produced alterations in observer search strategies, and it is these search strategy adjustments which lead to discrepancies within the obtained dimensional findings. A second possibility for the dimensional inconsistencies may be due to the differences in visual saliencies of both the local area surrounding the target (local-target-area), and the entire visual search field (global area). While local and global visual saliencies remained equal within Bertolotti and Strybel [7], Vu et al. [6] manipulated local and global visual saliencies via distractor density. Results show when the local-target-area surrounding the target was visually cued via global-local distractor densities, vertically displaced auditory cues produced significantly higher search times than horizontally displaced auditory cues. It appears that as the local target area becomes increasingly visually salient as a result of global-local distractor density combinations, facilitation of the localization stage of the visual search process may occur, accounting for the dimensional inconsistencies.

To examine these two possibilities, and to gain a clearer understanding of the inconsistent dimensional findings in the previous literature, the current study observed the effects of auditory spatial cue precision, target uncertainty, and local target area visual saliency on the visual search process.

2 Method

2.1 Participants

Twelve students (5 males, 7 females) with a mean age of 23.92 years (SD = 3.37 years) participated in the study. All participants were students from California State University, Long Beach, and reported normal hearing and normal to corrected-to-normal vision. Participants were paid $60 for completing the experiment.

2.2 Apparatus

The apparatus and materials utilized were identical to that of Bertolotti and Strybel [7]. Positioned at the center of a semi-anechoic room, covered by Markerfoam 10.16 cm acoustic foam sheets (absorption coefficients exceed .90 for frequencies greater than 250 Hz), was a large acoustically transparent projection screen. Figure 1 illustrates a sample of the visual search field consisting of a target (1° x 1° arrowhead pointing right or left), distractors (1° x 1° arrowheads pointing up or down), and a visual cue (described later). Contrast between the white targets and distractors and the black screen was held constant at roughly 75%, with local and global distractor-densities remaining constant at 32%. Two data projectors connected to a microprocessor located in an adjacent room were used for presenting the visual stimulus during the experiment.

For the auditory stimulus, Tucker-Davis Technologies' audio modules were used for generating and presenting the 65-dB A-weighted auditory stimulus. The auditory stimulus was a series of 300 ms broadband noise bursts separated by 100 ms quiet intervals that remained on until a response was made. Forty-five 7.6 cm Blaupunkt speakers mounted behind the screen in six circular centric rings produced the auditory stimuli. The positioning of the speakers, measured from the fixation point in the middle of the visual search field, created three distance ranges (12° to 18°; 24° to 30°; 36° to 42°) with midpoints of 15°, 27°, and 39°. The fixation point consisted of a 1° x 1° crosshair, located at the center of the visual search field.

In order to record participant responses, a four button response box was utilized. The two top buttons were used to identify the target (left button—arrowhead pointing to the left; right button—arrowhead pointing to the right), while either of the two bottom buttons were used to report no target present.

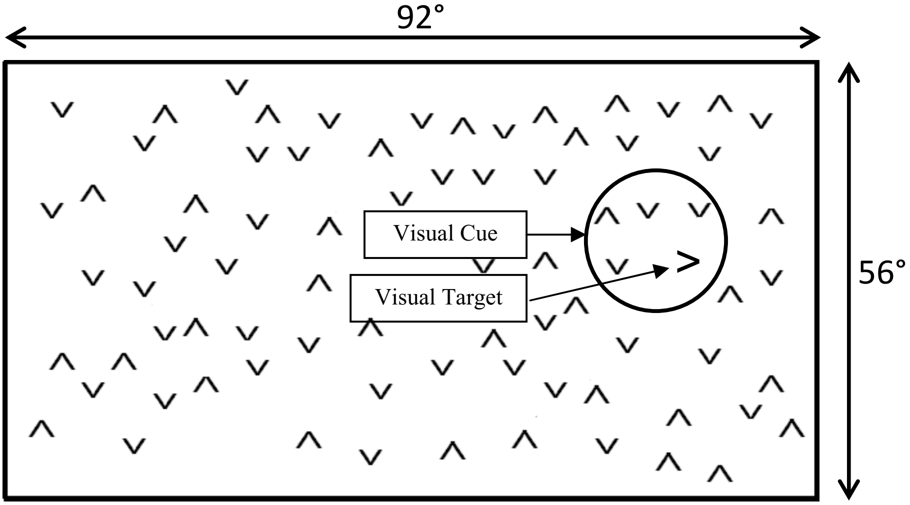


Fig. 1. Visual search field indicating visual cue (circle: 6.5° radius), and consisting of a visual target (arrowhead pointing right) and distractors (arrowheads pointing up or down)

2.3 Design

Four independent variables were manipulated: cue condition, auditory cue error, target percentage, and target distance. Cue condition was the first variable under investigation, and consisted of two levels: auditory cue and audiovisual cue. Within the auditory cue condition, the participant was provided with only an auditory spatial cue which signaled the location of the target. In the audiovisual cue condition, the participant was provided with a visual cue in addition to the auditory spatial cue. Following the local target area dimensions of previous research [8], the visual cue consisted of a circle (6.5° radius) identifying the local target area (Figure 1). Auditory cue error, the second variable under investigation, varied the direction and distance between the auditory spatial cue and the target. There were two displacement directions (horizontal and vertical), and three error magnitude distances (0°, 4°, & 8°).

Target percentage was the next independent variable, and consisted of three levels: 25%, 50%, and 100%. In the 25% target percentage condition, 25% of the trials within a session contained targets. The second target percentage condition provided targets on 50% of the trials within a given session. The final target percentage condition provided a target on 100% of the trials. For no target trials, a distractor (arrowhead pointing either up or down) was placed in the exact position where the target would have been.

The final independent variable was target distance from the point of fixation, and there were three distance ranges: 12° to 18°, 24° to 30°, and 36° to 42°. These three distance ranges created midpoints of 15°, 27°, and 39° accordingly. Again, a 1° x 1° crosshair located at the center of the visual search field was considered the point of fixation. It must be mentioned that at the farthest distance from the point of fixation (±39°), horizontal displacement of the auditory cue outside the local target area (8°)

was possible only in the direction towards the point of fixation due to the close proximity of the speakers to the peripheral edge of the visual search field. The dependent variable for the current study was target and no target search times.

2.4 Procedure

Participants were required to complete a screening form and instructed on the purpose of the study before signing the informed consent. Once completed, participants were seated 127 cm from the projected visual search field and instructed to located and identify whether a visual target was present or absent, and if present, identify the direction it was pointing. A practice session (20 trials) was given in order to familiarize the participant with the experiment and how to respond. After the participant reached a response accuracy rate of 95%, the participant was instructed that the experiment was about to begin, and was provided information regarding the visual cue and target percentage for the current block.

Upon the start of a trial, the participant was required to fixate on the crosshair located at the center of the visual search field containing many X's. After a period of 500 ms—1000 ms, the crosshair disappeared, the X's turned into distractors, and a target accompanied by the appropriate cue (auditory or audiovisual) was presented. The participant was instructed to scan the visual search field and report, via the response box, the direction of the target (left or right) or no target. After the response the trial terminated, the visual search field disappeared, and the crosshair positioned at the center of the search field accompanied by the X's reappeared to signify the start of the next trial. If no response was made after 8 seconds, the trial was terminated.

Participants completed a total of six 1-hour blocks, of which three blocks were completed with only an auditory cue and three blocks were completed with an audiovisual cue, in random order. Within each cue condition, the three blocks were further separated by target percentage, resulting in each cue condition containing a separate 1-hour block for each target percentage condition (25%, 50%, & 100%). Participants were informed of both cue condition and target percentage prior to starting a block. Within each 1-hour block, three separate 20-minute sessions were completed. Of the three sessions completed within each 1-hour block, the first session was considered practice while the remaining two were analyzed.

3 Results

To adjust for skewed distribution of search times, ANOVAs were run using log transformations of search times; however, actual search times are used in figures for ease of interpretation. Any violations of Mauchly's test of sphericity were corrected, and adjusted degrees of freedom from Huynh-Feldt estimates of sphericity were used. Post-hoc analyses were completed with the use of Tukey post-hoc test.

Separate ANOVAs were run for each cue condition due to the large differences in search times between the audiovisual cue condition (target present: M = 911.99 ms, SEM = 74.8 ms; target absent: M = 835.375 ms, SEM = 69.86 ms) and the auditory cue condition (target present: M = 1134.69 ms, SEM = 117.57ms; target absent:

M = 3748.03ms, SEM = 396.32ms). For the target present trials, two 3 (target percentage: 25%, 50%, & 100%) X 3 (target distance: 15°, 29°, & 39°) X 2 (error direction: horizontal and vertical) X 3 (error magnitude: 0°, 4°, & 8°) repeated measures analysis of variance (ANOVAs) were run on target present search times for each cue condition (auditory and audiovisual). For the target absent trials, two 2 (target percentage: 25% & 50%) X 3 (target distance: 15°, 27°, & 39°) X 2 (error direction: horizontal and vertical) X 3 (error magnitude: 0°, 4°, & 8°) repeated measures analysis of variance (ANOVAs) were run on target absent search times for each cue condition (auditory and audiovisual). ANOVAs run on target absent trials for each cue condition contained only two percentage conditions (25% & 50%), due to a lack of no target trials within the 100% target percentage condition. Response accuracy was 98.08% (*SD* = 1.14%), and trials which participants time-out were removed prior to analysis.

3.1 Auditory Cue Condition

For target present trials, a significant main effect of error direction was obtained (*F*(1, 11) = 10.425, *p* = .012). Search times with horizontally displaced cues (*M* = 1254.79 ms, *SEM* = 80.84 ms) were significantly higher than search times with vertically displaced cues (*M* = 1132.83 ms, *SEM* = 63.60 ms). A significant main effect of error magnitude also was obtained for target present trials (*F*(2, 22) = 15.121, *p* = .001), with search times significantly increasing with error magnitude: 0°: *M* = 1102.21 ms, *SEM* = 67.43 ms; 4°: *M* = 1165.88 ms, *SEM* = 63.55 ms; and 8°: *M* = 1313.34 ms, *SEM* = 87.98 ms. No significant main effects were found on target absent latencies within the auditory cue condition.

The significant main effects for target present trials were qualified by a significant error direction X error magnitude interaction (*F*(1.564, 17.487) = 6.623, *p* = .022), shown in Figure 2a. While search times for target present trials increased with error magnitude for both horizontal and vertical displacement, search times remained consistently higher when the auditory cue was displaced horizontally compared to vertically, particularly at 8° of error magnitude displacement.

Simple effects analysis revealed a significant simple effect of error direction at both 4° (*F*(1, 11) = 12.348, *p* = .005) and 8° (*F*(1, 11) = 10.407, *p* = .008). Horizontal displacement was found to be a significantly greater detriment to target identification than vertical displacement, regardless of error magnitude. To examine the linear relationship between both error directions, as well as the increasing rate of search time, mean search times were regressed against error magnitude. Only horizontal displacement was found to explain a significant portion of the variance in target identification search times (horizontal displacement: R^2 = .97; *F*(1, 11) = 36.403, *p* = .01; vertical displacement: R^2 = .78; (*F*(1, 11) = 3.581, *p* = .30). Furthermore, a slope of 43.57 (±7.22) was obtained for horizontal displacement, and 9.213 (±4.87) for vertical displacement. Thus, for every 1° increase in horizontal displacement, search times increased by approximately 43 ms, and every 1° increase in vertical displacement, search times increased by approximately 9 ms.

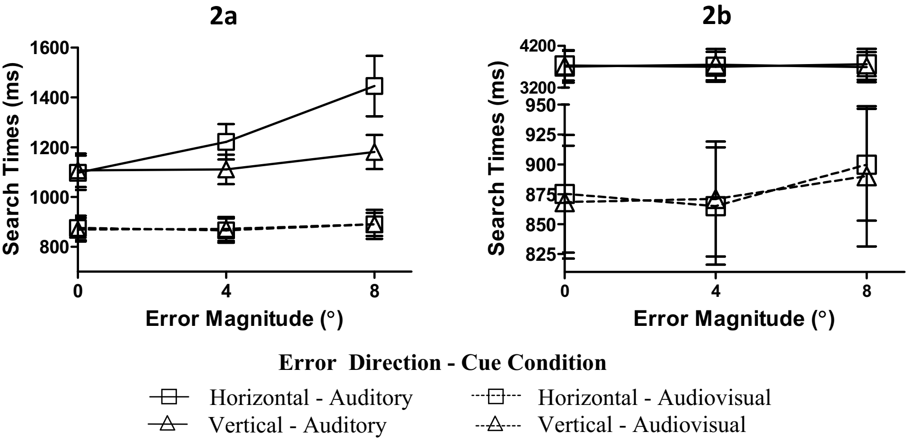


Fig. 2. Search times for auditory and audiovisual cue conditions as a function error direction and error magnitude: a) target present trials b) target absent trials

A significant error magnitude X target percentage interaction also was obtained for only the target present trials (F(4,20) = 2.98, p < .05). As illustrated in Figure 3, at 25% target percentage, error magnitude had little effect on search times; however, as target percentage increased, search times increased with error magnitude. Simple effect of error magnitude on target percentage demonstrated a significant simple effect at 50% (*F*(2, 22) = 21.014, *p* < .001) and 100% (*F*(1.491, 16.4) = 22.534, *p* < .001). Post-hoc tests found a significant increase in search times at 50% as error magnitude increased from 0° to 8°and 4° to 8°, and a significant increase in search times at 100% for every error magnitude increment. Thus, while error magnitude did not affect target present trial search times at 25%, search times significantly increased for every error magnitude increment at both 50% and 100%, with the exception of the 0° to 8° increment at 50%.

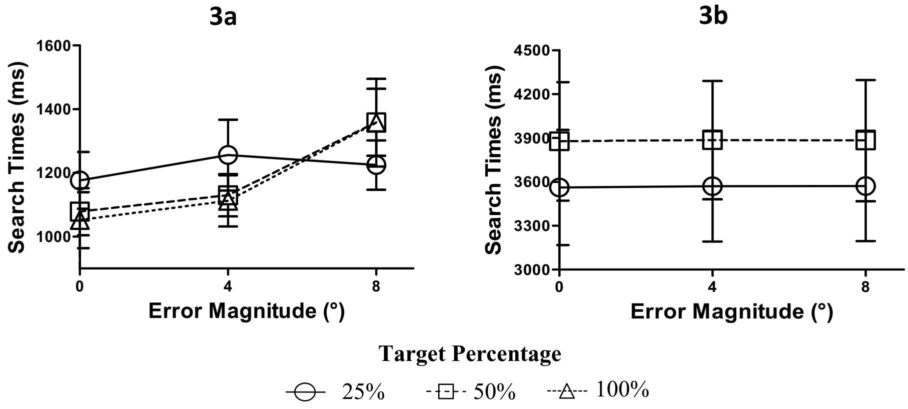


Fig. 3. Search times for the auditory cue condition as a function target percentage and error magnitude: a) target present trials b) target absent trials (contains only 25% and 50% conditions)

3.2 Audiovisual Cue Condition

For target present trials, a significant main effect of target percentage was obtained (*F*(2, 11) = 10.03, *p* = .002). While search times decreased as target percentage increased (25%: *M* = 921.40 ms; *SEM* = 70.32 ms; 50%: *M* = 869.09 ms, *SEM* = 57.19 ms; 100%: *M* = 843.99 ms, *SEM* = 56.58 ms), post-hoc testing found only one significant difference in search times, 25% vs. 100%. A significant main effect of target distance was also found for target present trials (*F*(2, 11) = 19.539, *p* < .001). Post-hoc testing determined that as target distance increased, search times significantly increased only at target distances between 15° (*M* = 871.26 ms, *SEM* = 62.98 ms) vs. 39° (*M* = 902.01 ms, *SEM* = 58.47 ms), and 27° (*M* = 861.21 ms, *SEM* = 57.19 ms) vs. 39°. No significant interactions were obtained for target present trials within the audiovisual cue condition.

For target absent trials, a significant main effect of target percentage was obtained (*F*(1, 11) = 5.951, *p* = .033), with search times significantly increasing with target percentage. A significant main effect of target distance was also found for target absent trials (*F*(2, 22) = 32.624, *p* < .001). Significant increases in search times were found between 15° vs. 39, and 27° vs. 39°. However, between 15° vs. 27°, search times at 27° were significantly lower than search times at 15°. A significant main effect of error magnitude was found for target absent trials (*F*(2, 22) = 4.076, *p* = .031). Significant increases in search times were found between 0° to 4°, and 4° to 8°. No difference in search times were found as error magnitude increased from 0° to 8°.

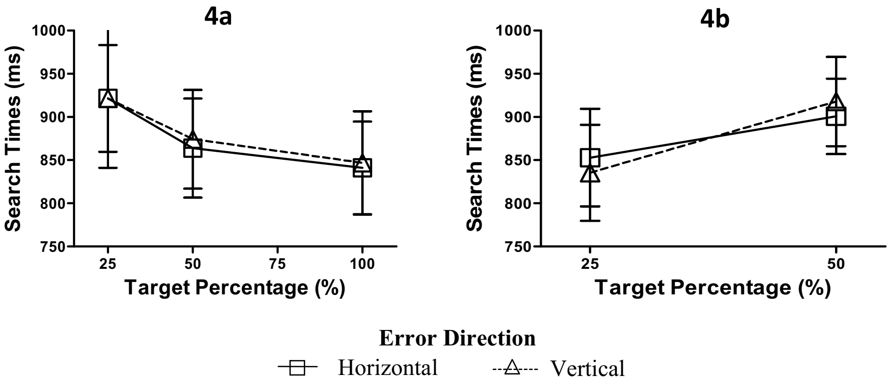


Fig. 4. Search times for the audiovisual cue condition as a function error direction and target percentage: a) target present trials b) target absent trials

For target absent trials, a significant target percentage X error direction interaction was obtained (target present: *F*(2, 22) = .506, *p* = .614; target absent: *F*(1, 11) = 5.07, *p* = .046). As Figure 4 illustrates, search times increased with target percentage at both horizontal and vertical displacement. Simple effects of target percentage for error direction demonstrated a marginally significant simple effect of target percentage for horizontal displacement (*F*(1, 11) = 3.642, *p* = .083), and a significant main effect of target percentage for vertical displacement (*F*(1, 11) = 7.876, *p* = .017).

Search times increased with target percentage at both horizontal and vertical displacement, but the rate of increase was higher with vertical displacement. Furthermore, while horizontal displacement was a greater detriment to search times at 25%, vertical displacement was a greater detriment to search times at 50%. Simple effects of error direction for target percentage were examined, and a significant simple effect of error direction was obtained only at 25% ($F(1, 11) = 8.699, p = .013$). Horizontal displacement was a significantly greater detriment to search times than vertical displacement. For the target present trials (Figure 4a), while search times decreased with target percentage at both horizontal and vertical displacement, these were non-significant.

4 Discussion

Examining the dimensional findings within the current study, the results appear to be consistent with Bertolotti and Strybel [7]: horizontal displacement of an auditory cue was a greater detriment to the identification of a visual target than vertical displacement. Furthermore, within both the current and previous research [7], horizontally displacing an auditory cue either 4° or 8° from a visual target produced higher search times than vertically displacing an auditory cue either 4° or 8° from a visual target. Thus, an auditory cue displaced horizontally within the local-target-area increased search times more than a vertically displaced auditory cue outside the local-target-area. Given that this interaction did not depend on target percentage or target distance, it may reflect differences in localization between the horizontal and vertical plane; specifically, higher auditory spatial acuity in the horizontal plane compared to the vertical plane [9].

As for the first possibility of introducing target uncertainty, the current study found that when target uncertainty was introduced, horizontal displacement of an auditory cue continued to be a greater detriment to visual target identification than vertical displacement, regardless of error magnitude. Moreover, horizontal displacement was a greater detriment to target identification when target percentage was 100%, which suggests that target uncertainty may not be responsible for the incongruent directional displacement findings. In addition, within both the current study and Bertolotti and Strybel [7], search times significantly increased as the auditory cue error and target percentage increased. Therefore, it is suggested that search strategies for a visual target assisted by only an auditory cue are affected by target uncertainty in that as the likelihood of a target increased, greater dependence on the auditory cue for visual target identification was shown, which lead to a greater effect of auditory cue directional displacement. For the directional displacement when no visual target was present, the current study found an effect of directional displacement only as both target percentage and error magnitude increased.

With regards to the second possibility, the findings from the current study are consistent with the previous literature in that increasing local-target-area visual saliency significantly improves the visual search of a target [6] [8]. When provided with a reliable local-target-area visual cue and a variably displaced auditory cue, search times decreased by approximately 25% compared to search times with a variable displaced auditory cue alone. However, while Vu et al. [6] found variable vertical displacement of the auditory cue outside the local-target-area to be more of a detriment to the visual search process when the local-target-area was visually salient, the current study found the no effect of auditory cue displacement when the local-target-area was visually cued. Regardless of the error magnitude or directional displacement of the auditory cue, providing a visually salient local-target-area significantly improved visual target identification. It must be noted that while Vu et al. [6] varied global-local distractor-densities to manipulate the visual saliency of the local-target-area, the saliency of the local-target-area within the current study was varied by either providing a visual cue which accurately and consistently cued the local-target-area, or providing no visual cue. Possibly the high saliency of the local-target-area visual cue within the current study rendered the directional displacement of the auditory cue ineffective, hindering the ability to reconcile the incongruent findings within Bertolotti and Strybel [7] and Vu et al. [6].

One possible explanation for the ineffectiveness of auditory cue precision with audiovisual cues is that participants used the unreliable auditory cue to identify the hemi-field in which the target was located, while using the visual cue to determine the local-target-area and target location. This is suggested since providing a visually cued local-target-area was found to not only significantly improve target identification [6] [8], but also rendered the effect of auditory cue displacement less effective. Additional evidence for this possibility is that search times for targets assisted by an audiovisual cue were found to increase as target distance increased. Thus, it is suggested that while the auditory cue provided the direction in which the target was located compared to the center of the visual search field, the visual cue provide the local-target-area visual saliency needed to identify the local-target-area, and the target itself.

In summary, these results suggest that the effectiveness of audio spatial cueing systems depend more on precision cueing of horizontal target position compared with vertical target position because of the greater detriment in search latencies found with horizontally displaced cues. Moreover, with production cost much lower for auditory cues required for localization in the horizontal plane (ILDs and ITDs) compared to auditory cues required for localization in the vertical plane (spectral shape cues) [10], minimizing horizontal error would not only improve operator performance, but also would reduce development costs. The current study also has shown that providing an audiovisual cue can significantly improve target search latency compared to only an auditory cue. Furthermore, the benefits of providing an audiovisual cue were greater when participants were required to report no target, with performance improving as much as 300%. Thus, for environments in which false alarms are a possibility, audiovisual cues improve performance over an auditory cue alone.

Acknowledgements. This paper was partially supported by the Center for Human Factors in Advanced Aeronautics Technologies, a NASA University Research Center (NASA cooperative agreement NNX09A66A.)

References

1. Perrott, D.R., Cinseros, J., McKinley, R.L., D'Angelo, W.R.: Aurally aided visual search under virtual free-listening conditions. Human Factors 38, 702–715 (1996)
2. Perrott, D.R., Saberi, K., Brown, K., Strybel, T.Z.: Auditory psychomotor coordination and visual search. Perception and Psychophysics 48, 214–226 (1990)
3. Begault, D.R.: Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation. Human Factors 35, 707–717 (1993)
4. McKinley, R.L., D'Angelo, W.R., Hass, M.W., Perrott, D.R., Nelson, W.T., Hettinger, L.J., Brickman, B.J.: An initial study of the effects of 3-dimensional auditory cueing on visual target detection. In: Proceedings of the Human Factors and Ergonomics Society, vol. 39, pp. 119–123 (1995)
5. McKinley, R.L., Erickson, M.A.: Flight demonstration of a 3-D auditory display. In: Gilkey, G.H., Anderson, T.R. (eds.) Binaural and Spatial Hearing in Rea; and Virtual Environments, pp. 683–699. Erlbaum, Mahwah, NJ (1997)
6. Vu, K.L., Strybel, T.Z., Proctor, R.D.: Effects of displacement magnitude and direction of auditory cues on auditory spatial facilitation of visual search. Human Factors 49(3), 587–599 (2006)
7. Bertolotti, H., Strybel, T.Z.: he effects of auditory cue precision on target detection and identification. Presentation session presented at the meeting Human-Computer Interaction Institution International, Las Veges, NV (2005)
8. Rudmann, D.R., Strybel, T.Z.: Auditory spatial cueing in visual search performance: Effect of local vs. global distractor density. Human Factors 41, 146–160 (1999)
9. Strybel, T.Z., Fujimoto, K.: Minimum audible angles in the horizontal and vertical planes: Effects of stimulus onset asynchrony and burst duration. Journal Acoustical Society of America 108, 3092–3095 (2000)
10. Shinn-Cunningham, B.G.: Spatial auditory displays. In: Karwowski, W. (ed.) International Encyclopedia of Ergonomics and Human Factors, 2nd edn. Taylor and Francis, Ltd., Abington (2002)

Interpretation of Metaphors with Perceptual Features Using WordNet

Rini Bhatt, Amitash Ojha, and Bipin Indurkhya

Cognitive Science lab, International Institute of Information,
Technology Hyderabad- 500 032, India
rini_bhatt@students.iiit.ac.in, {amitash.ojha,bipin}@iiit.ac.in

Abstract. Metaphors based on perceptual similarity play a key role in stimulating creativity. Here, we present a metaphor interpretation tool using features of source and target to generate perceptual metaphors which might be conceptually very different, thereby generating new interpretations from familiar concepts.

Keywords: Perceptual metaphors, conceptual combination, creative cognition, juxtaposition, conceptual association, metaphorical interpretation.

1 Introduction

“What is creativity?” This question has no universally agreed answer. [1] For our purpose, however we define creativity as the ability to generate ideas or artifacts that are novel, surprising and valuable, interesting, useful, funny, beautiful, etc. According to Perkins “Creativity is not a special 'faculty', nor a psychological property confined to a tiny elite. Rather, it is a feature of human intelligence in general”. [2] It rests on everyday capacities such as the association of ideas, analogical thinking, searching a structured problem-space, and reflective self-criticism. Various kinds of creativity have been mentioned in the literature. But broadly there are three ways in which processes can generate new ideas: *Combinational, exploratory* and *transformational.* [3]

We are concerned with combinational creativity, which is the production of novel (unfamiliar, improbable) combinations of familiar ideas. It has been studied in AI by the many models of analogy and by the occasional joke generating programs or database metaphor generation tools. One such system is JAPE, which models the associative processes required to generate punning jokes. Such processes are far from random, and depend on several types of knowledge such as lexical, semantic, phonetic, orthographic, and syntactic [4]. For analogy, most AI models generate and evaluate analogies by exploiting the programmer's careful pre-structuring of the relevant concepts. This guarantees that their similarity is represented, and makes it likely that the similarity will be found by the program [5].

In this paper we present a metaphor interpretation tool using perceptual features of Source concept and Target concept. The tool can help users to generate perceptual metaphors by evoking their ability to make free associations. The tool, in other words, assists users to create unfamiliar association from familiar concepts.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 21–27, 2011.

1.1 Pictorial Metaphors

Metaphorical thinking is known to play a key role in stimulating creativity. In a metaphor, one kind of object or idea is used in place of another to suggest some kind of likeness between them. They serve in making connections between things that are not usually seen as connected in any conventional way. For example, computer science instructors often explain the function of the Control Unit in a computer's Central Processing Unit by saying it is 'the traffic policeman of the computer'. The general use of the term 'memory' to denote computer storage is metaphoric. Thus, metaphors are instruments of divergent processes because they synthesize disparate ideas.

Similarly, in pictorial metaphors two concepts are juxtaposed or replaced to create a unified figure suggesting one concept being described in terms of another. The perceived incongruity in the image invites the viewer to interpret the image metaphorically. Though metaphors have mostly been studied as a literary device, but research on pictorial metaphor has over the past 25 years yielded a few theoretical studies [6], [7], [8], [9].

1.2 Perceptual Similarities in Pictorial Metaphors

We have hypothesized that perceptual similarity between two images at the level of color, shape, texture, etc. helps to create metaphorical associations [10]. Elsewhere we have shown that participants have preference for perceptually similar images in generating metaphorical interpretations. Also they help in creating more conceptual associations between Source and Target [11].

1.3 Computers and Creativity

For creativity a cognitive agent needs to break conventional conceptual associations and this task is difficult for human beings because we inherit and learn, in our lifetime, to see the world through associations of our concepts. It requires a significant amount of cognitive effort to break away from these associations. Computers, on the other hand, do not have such conceptual associations. Therefore it must be easier for the computers to break away from these conceptual association simply because they do not have them to begin with. It follows that computers are naturally predisposed towards incorporating creativity [12].

2 Using WordNet for Generating Metaphorical Interpretation

2.1 Idea and Aim

The present system aims to extract perceptual features of Source concept and Target concept and uses it to anchor conceptual associations between them for a metaphorical interpretation. The larger goal of the system is to be able to generate metaphorical interpretations for an image by extracting perceptual features (for visual: color, shape, texture, orientation, etc, for touch: Hot, cold, rough, smooth, etc.). With the availability of these features, various concepts can be evoked (which are usually not evoked just by the conceptual description of the object) and as a result distant associations can be generated which may or may not be creative. But as a matter of fact the perception of the creativity lies in the interaction between cognitive agent and stimuli. So, the system can suggest some combinations (that are guided by the perceptual features of concepts and are not arbitrary) using WordNet, which may be seen creative by users.

2.2 WordNet

WordNet is a large lexical database of English, developed under the direction of George A. Miller (Emeritus). Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are Interlinked by means of conceptual-semantic and lexical relations.

It can be visualized as a graph with edges between words with some relationship (synonym, antonym, part-of, etc.) between them. Such words will, henceforth, be referred to as neighbors bearing the graph analogy in mind. WordNet 2.0 for Linux allows us to feed a word and the relationship (synonym - noun/adjective,etc.) on the commandline and outputs the words which satisfy this relationship - the ‘neighbors’ of that word. This is exploited while generating metaphors of multi-word sets of perceptual and conceptual features.

2.3 WordNet and Metaphor Generation

There have been various attempts to use WordNet for generating metaphors. [13],[14]. [15]. We take a different approach and try to use perceptual features of concepts to anchor the metaphorical interpretation. The creativity assistance tool generates metaphors for any arbitrary combination of words divided into 2 categories “Perceptual” and “Conceptual”. The generated ‘metaphors’ depend upon certain parameters, for instance, it takes into account, a certain threshold within which to look for neighbors of a word. It also allows for semantic constraint on the metaphors generated. As mentioned, with WordNet, we can find synonyms, antonyms, hypernyms, etc. of a given word. In our tool, we have considered the three mentioned here and any one type or a combination of these in any order can be used to generate interpretations by taking that particular path.

2.4 Threshold

It is important to define this threshold as WordNet is organized as a connected graphand the search will go on infinitely unless a certain threshold value for the number of levels which can be visited for finding relevant words, is specified. Moreover, a word which is related to the given word but is, say, thirty levels away will likely be semantically and relationally dissimilar and therefore, will not satisfy our primary objective of finding different interpretations for the given set of features.

2.5 Anchoring

Since various features are to be considered in the perceptual set as well as the conceptual set, this entire set of features must be taken as a whole to identify interpretations which are relevant to the entire picture instead of treating the features separately. In order to do this, there is a need to identify ‘an anchor’ word which is closest to the set of all features and serves as the link between them.

This anchoring was done by finding the words up till the mentioned threshold for all the input features and then finding the words common to all the input features,i.e. if the input words (perceptual features) are ‘big’ and ‘red’ then the tool discovers the neighbors till the specified threshold and now from those two lists finds the closest common word.

Once the common word or the ‘anchor’ has been found for both the perceptual features and the conceptual features, these anchor words are used to generate word-pairs. Neighbors of the anchor words are discovered till the specified threshold and these separate interpretations of the perceptual and conceptual anchor are then combined to give P*C two-word phrases as metaphors, where

P = number of interpretations of the perceptual anchor word
C = number of interpretations of the conceptual anchor word.

The user could specify search till 2 levels for relation hypernym and then 3 levels for relation synonym. So here, first hypernyms would be generated till 2 levels and for the words derived, synonyms would be generated till 3 levels.The levels of semantic relations and their order bears significance on the results generated, i.e. whether five levels of hypernym are explored followed by two levels of synonyms or vice versa.

"The heuristic of choosing the first listed sense in a dictionary is often hard to beat, especially by systems that do not exploit hand-tagged training data." [16]. Instead of following a particular method of disambiguation, we have chosen to consider both noun and adjective for perceptual features while only nouns have been considered for conceptual features. This is to allow generation of metaphors from both senses of the word - choosing to limit it to one might detract from the creativity of metaphor generation. Choosing to include nouns while discovering neighbors in case of perceptual features includes a trade-off - we get a wider range of interpretations now which include some that would be have been absent if the word was considered as only noun or only adjective, however, the list would also include some interpretations with a bad similarity measure.

3 Testing and Results

We present an example which illustrates one of the results. In this example, 'red' and 'hot' are entered as the perceptual features. Now the tool seeks to find an 'anchor' for the given perceptual features within the specified threshold (in this example it is 2 levels). Bearing the graph structure in mind, first all the immediate neighbors (of the form synonym-noun and synonym-adjective) of the ‘red’ and the word 'hot' are discovered. This is level one i.e. these words are at distance '1' from the original words entered. From these words, a further level of words is discovered, i.e. the immediate neighbors of level 1 words. From the two lists - one for 'red' and the other for 'hot', the common words are determined. For these words, a common measure of distance from both words is computed. This is calculated as the vector distance between the words 'red' and 'hot' from the common word.

The one with the least such distance qualifies as the anchor word. This graph is a partial representation of the discovery mechanism followed to find the anchor word for perceptual features. After computing the two lists, 'wild' is found as the common word at minimum distance - 2 - as it is a layer 2 word for both 'red' and 'hot'. The words encircled in red color trace the path from 'red' and 'hot' to 'wild'. Thus, 'violent' is a synonym of 'red' and 'wild' is a synonym of 'violent' thereby giving a path from 'red' to 'wild'. Similarly for 'hot', 'raging' and 'wild'. Thus, 'wild' becomes the anchor word for this set of perceptual features. A similar mechanism is followed for conceptual set and a common word is found at the similar level (or at user defined level). Now, once the system has one common word for perceptual features and one common word for given conceptual features, same procedure is followed and list of neighboring words are generated as result for a certain level.

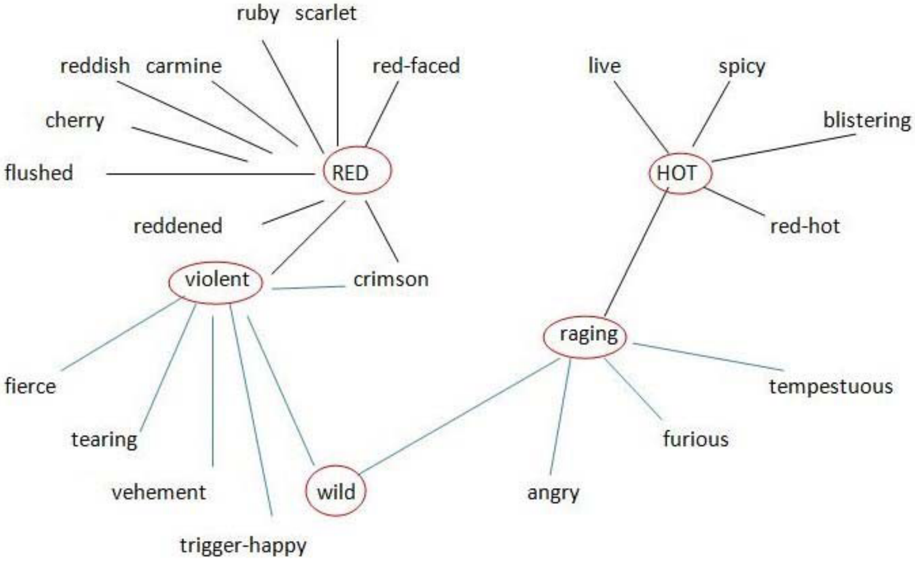


Fig. 1. Graphic representation of an example to find a common word for perceptual features

Further, A user testing was done with some of the results. 10 sets of word-pairs (each containing 50 word-pairs generated by system) were taken as test stimuli. These pairs were produced with 2 levels of synonym and 1 level of hypernym. To generate stimuli two perceptual features and one conceptual feature was used. 7 participants were asked to categories these pairs as 1. Metaphor, 2. Anomaly and 3. Literal. We found that 63 percent of word-pairs were categorized as “metaphor”, 32 percent of pairs were categorized as “anomaly” and 5 percent of word-pairs were categorized as “literal”. The difference between them was statistically significant. F (2, 12) =4.14 p < .05. (Figure 2)

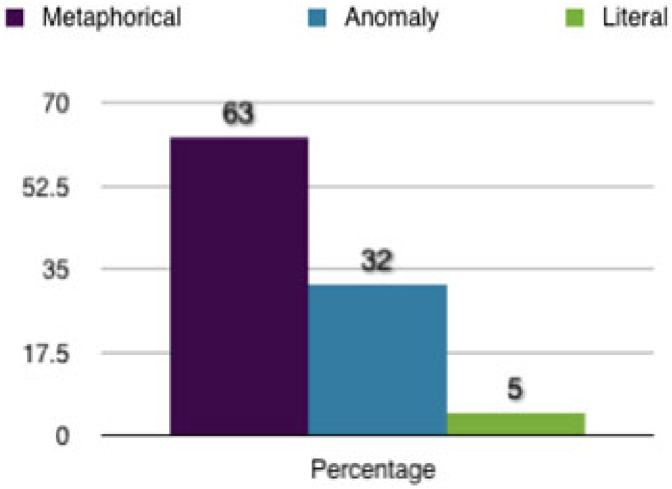


Fig. 2. Categorization of generated word-pairs

4 Limitations and Future Work

The system is in its preliminary stage and has to be developed and tested further. At present it has various limitations: In the system, selection of levels is completely arbitrary and there is no fixed threshold to decide the metaphoricity of word-pairs. Some of the results at the very lower level can be interpreted metaphorically and some can not be interpreted metaphorically at higher level. We plan to solve this problem by measuring the distance between source and target in some conventional metaphors. An average distance can guide us to suggest that at what level chances are high for two concepts to be related metaphorically.

The system aims to extract perceptual features automatically by a given picture using image search engines [17]. But for this users will be asked to tag the object depicted in the image. Image search engine will identify the color of the background, shape of the object, etc and then provide set of perceptual features. These perceptual features combined with conceptual feature/s can be used for further interpretation task.

References

1. Sternberg, R.J. (ed.): Handbook of creativity. Cambridge University Press, Cambridge (1999)
2. Perkins, D.N.: The mind's best work. Harward university press, Cambridge, MA (1981)
3. Boden, M.A.: What is creativity? In: Boden, M.A. (ed.) Dimensions of creativity, pp. 75–118. MIT Press, Cambridge (1994)
4. Binsted, K., Ritchie, G.: An implemented model of punning riddles. In: Proceedings of the Twelfth National Conference on Artificial Intelligence, Seattle, pp. 633–638 (1994)

5. Forbus, K.D., Gentner, D., Law, K.: MAC/FAC: A model of similarity based retrieval. Cognitive Science, 141–205 (1994)
6. Kennedy, J.M.: Metaphor in Pictures. Perception 11, 589–605 (1982)
7. Forceville, C.: Pictorial Metaphor in Advertising. Routledge, London and New York (1996)
8. Whittock, T.: Metaphor and Film. Cambridge University Press, Cambridge (1990)
9. Caroll, N.: Visual Metaphor. In: Hintikka, J. (ed.) Aspects of Metaphor, pp. 189–218. Kluwer Academic Publishers, Dordrecht (1994)
10. Indurkhya, B.: Metaphor and cognition. Kluwer Academic Publishers, Dordrecht (1992)
11. Ojha, A., Indurkhya, B.: Role of perceptual metaphors in metaphorical comprehension. In: The Proceedings of European Conference on Visual Perception, Regensburg, Germany (2009)
12. Indurkhya, B., Kattalay, K., Ojha, A., Tandon, P.: Experiments with a creativity-support system based on perceptual similarity. In: Proceedings of the 7th International Conference of Software Methodologies, Tools and Techniques; Encoding information on metaphoric expressions in WordNet like resources. In: Proceedings of the ACL 2003 workshop on lexicon and figurative language, vol. 14, pp. 11-17 (2003)
13. McCarthy, D., Keeling, R., Weeds, J.: Ranking WordNet senses automatically. Cognitive Science Research Paper, 569
14. Tandon, P., Nigam, P., Pudi, V., Jawahar, C.V.: FISH: A Practical System for Fast Interactive Image Search in Huge Databases. In: Proceedings of 7th ACM International Conference on Image and Video Retrieval (CIVR 2008), Niagara Falls, Canada (2008)

Acoustic Correlates of Deceptive Speech – An Exploratory Study

David M. Howard and Christin Kirchhübel

Audio Laboratory, Department of Electronics, University of York, UK
{dh,ck531}@ohm.york.ac.uk

Abstract. The current work sets out to enhance our knowledge of changes or lack of changes in the speech signal when people are being deceptive. In particular, the study attempted to investigate the appropriateness of using speech cues in detecting deception. Truthful, deceptive and control speech was elicited from five speakers during an interview setting. The data was subjected to acoustic analysis and results are presented on a range of speech parameters including fundamental frequency (f0), overall intensity and mean vowel formants F1, F2 and F3. A significant correlation could not be established for any of the acoustic features examined. Directions for future work are highlighted.

Keywords: Deception, speech, acoustic, Voice Stress Analyzer.

1 Introduction

It is acknowledged that information can be gained about a human speaker from the speech signal alone. Possible knowledge that can be derived include a speaker's age, regional and social background, the presence of speech or voice based pathology, voice/language disguise, speaking style, and influence of alcohol intoxication.

The voice can also give information about a speaker's affective state. Listening to a third party conversation, lay listeners are usually able to tell whether the speakers are happy, sad, angry or bored. Whilst at an interpersonal level it is possible to perceive accurately emotional states, empirical research has not been successful in identifying the speech characteristics that distinguish the different emotions. Compared to the investigation of speech and emotion, research into psychological stress has been somewhat more successful in establishing the acoustic and phonetic changes involved [1]. However, facing similar methodological and conceptual obstacles, it is more appropriate to refer to the correlations that have been discovered so far as 'acoustic tendencies' rather than 'reliable acoustic indicators'.

If it is possible to deduce a speaker's emotional condition from listening to their voice, could it also be viable to make judgements about their sincerity from speech as well? For centuries people have been interested and fascinated by the phenomenon of deception and its detection. Indeed, a device that would reliably and consistently differentiate between truths and lies would be of great significance to law enforcement- and security agencies [2]. In recent times, claims have been brought forward involving voice stress analysers (VSA) which are said to measure speaker

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 28–37, 2011.

veracity based on the speech signal. Scientific reliability testing of these products has mainly resulted in negative evaluations [3,4,5]. While testing of these products is a necessary part of their evaluation, it is believed that a more fundamental step has been overlooked. Prior to examining the reliability of a test it should be ascertained whether the assumptions on which the test is based are valid [4]. In other words, it needs to be established whether a relationship exists between deception, truth and speech, and if so, what the nature of this relationship is.

Surprisingly, very little research has been carried out on the acoustic and phonetic characteristics of deceptive speech. There are a number of studies [6,7,8] that have analysed temporal features such as speaking rate, pauses, hesitations and speech errors but only a few studies have investigated frequency based parameters as, for instance, mean f_0 and f_0 variability [9,10]. Evidence for the analysis of vowel formants and voice quality in connection with deceptive speech is rare. Recently completed work by Enos [11] is one of the first attempts to analyse deceptive speech using spoken language processing techniques. This work provides a basis in this complex area but more research is needed within this subject matter in order to improve our understanding of deceptive speech and consequently, to assess whether differentiating truthfulness and deception from speech is a realistic and reasonable aspiration.

2 Method

2.1 Participants

The data consists of an opportunistic sample of five male native British English speakers between the ages of 20 to 30 years (mean age = 24.7 years; SD = 3.65 years). The majority were drawn from the student population at the University of York and were from the northern part of England. None of the speakers had any voice, speech or hearing disorders.

2.2 Experiment Procedure

The procedure was modified from Porter [12] and is based on a mock-theft paradigm. The experiment was advertised as being part of a security research project and participants received £5 for participating with the chance of earning more money through the trial. Having arrived at the experimental setting participants were told that the University was looking to implement a new security campaign in order to reduce small scale criminal activity (e.g. theft on campus). The security scheme would involve employing non-uniformed security wardens who:

a) Patrol in selected buildings
b) Perform spot checks on people
c) Interview students suspected of having been involved in a transgression

The participants were then led to believe that the researchers were testing the effectiveness of this security system and, in particular, the extent that wardens would be able to differentiate between guilty and non-guilty suspects. Further to this the volunteers were informed that the experiment was also part of a communication investigation and therefore audio data would be collected. Having given written consent that they are prepared to continue with the study participants had to complete three tasks:

Task 1: sitting in a quiet office room ('preparation room'), participants were asked to complete demographic details. On completion of the forms they were taken to an interview room where they were involved in a brief conversation which formed the baseline/control data.

Task 2: participants were provided with a key and directions. They were asked to go into an office and take a £10 note out of a wallet located in a desk drawer and hide it on their body. They were advised to be careful in order not to raise suspicion or to draw attention of the security warden who was said to be in the building and who might perform a spot check.

Task 3: a security interview was conducted in which the mock security warden questioned the participant about two thefts that had allegedly occurred in the previous hour. The participant committed one of the thefts (theft of £10 note) but not the other (theft of digital camera from the 'preparation room'). Participants were required to convince the interviewer that they were not guilty of either theft. With respect to the camera theft, participants could tell the truth but when the interviewer asked about the £10 note the participant had to fabricate a false alibi. Each participant had 10 minutes prior to the interview to formulate a convincing story.

If the participants were successful in convincing the interviewer that they did not take either the camera or the £10 they could earn an extra £5 in addition to their basic £5 participation payment. If they failed on either however, they would lose the extra payment and be asked to write a report about what had happened.

2.3 Recording Setting/Equipment

The experiment was conducted in the Linguistics department at the University of York. A vacant office room was used as the 'preparation room' in which participants completed task 1 and prepared for task 3. The 'target room', an unoccupied office room, from which the £10 was taken, was situated at the other end of the corridor approximately 200m away from the 'preparation room'. The baseline/control data and the security interview were recorded in a small recording studio. The interviewer and participants were sat down, oriented at approximately 90 degrees to each other. To ensure that the distance between microphone and speaker was kept constant an omnidirectional head-worn microphone of the type DPA 4066 was used. The microphone was coupled to a Zoom H4 recorder.

2.4 Parameters Analysed

The experiment took the form of a within-subjects design. Truthful and deceptive speech was elicited from participants during the interviews. In addition baseline (control) data was recorded prior to the interviews. The three speaking conditions will be referred to as Baseline (B), Truth (T) and Deception (D) in this article.

A number of speech parameters was analysed including mean fundamental frequency (f_0 mean) and standard deviation (f_0 SD). The changes in mean energy across speaking conditions were computed as well as mean vowel formants F1, F2 and F3. Every speaker provided one file for each of the three speaking conditions, resulting in 60 files for analysis. The duration of the files was between about 3-5 minutes.

2.5 Measurement Equipment/Technique

'Sony Sound Forge' software was used for initial editing of the speech files. The acoustic analysis was performed using 'Praat' speech analysis software [13].

The f_0 based parameters were measured on the previously edited files using a Praat script developed by Philip Harrison.

In order to measure intensity, the files were edited in Praat so as to only contain speech (i.e. all silences were removed). The mean energy was then determined using a function of the Praat software. Rather than expressing the intensity values in absolute form, the differences between Baseline, Truth and Deception are reported in this paper.

Vowel formant measurements were extracted from Linear Predictive Coding (LPC) spectra using Praat's inbuilt formant tracker. The mean F1, F2 and F3 values were taken from an average of 10-20 milliseconds near the centre of each vowel portion. Any errors resulting from the in-built formant tracker were corrected by hand. To be counted in the analysis the vowels had to show relatively steady formants and be in stressed positions. For all speakers, 8 vowel categories[1]-FLEECE, KIT, DRESS, TRAP, NURSE, STRUT, LOT, and NORTH- were measured with one to 15 tokens (average 10 tokens) per category for each condition yielding a total of around 1200 measurements.

3 Results

The following section presents result for all 5 speakers. Statistical tests (T-tests) were employed where possible to assess the significance of the difference between the three conditions.

3.1 Fundamental Frequency (f_0)

Based on the f_0 mean and f_0 SD values obtained for each speaker and each condition, bar graphs were generated (Figure 1).

[1] Standard Lexical Sets for English developed by John C. Wells [14]. There are 24 lexical sets which represent words of the English language based on their pronunciation in Received Pronunciation (RP).

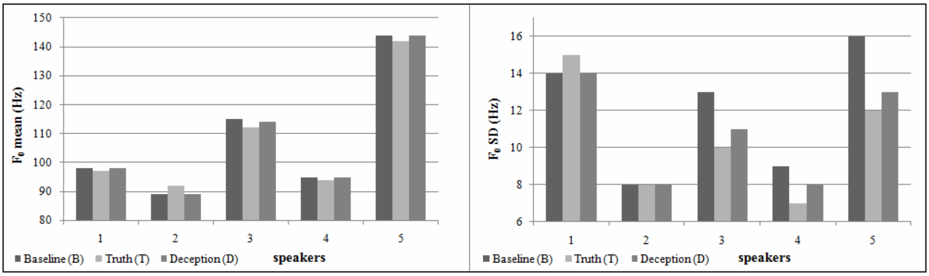


Fig. 1. f_0 mean (left) and f_0 SD (right) in Hz for all three speaking conditions for every speaker

Looking at Figure 1 (left) we can immediately see that there is not a great amount of difference in f_0 mean across conditions. There is a tendency that the f_0 mean of T is slightly lower than the values for B and D, which are close, but overall the mean f_0 values of all conditions are essentially similar. Examining the f_0 SD measures illustrated in the right hand side of Figure 1, we can perceive an overall trend in that there is less variation in f_0 in the Truth and Deception condition compared to the Baseline. The values of T and D for each speaker, in contrast, are rather comparable.

3.2 Intensity

Mean energy for each speaker is represented in Figure 2. No specific patterns can be generalized from the results. There is variability in direction and extent of change across speakers for both Truth and Deception. Apart from speakers 1, 3 and 5 the changes in mean energy are very small and interestingly, these three speakers show a uniform change in direction for both Truth and Deception.

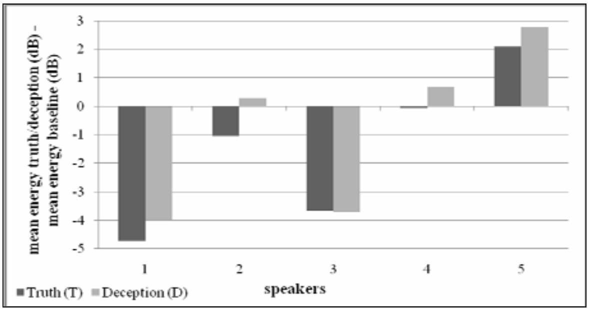


Fig. 2. Overall mean energy changes between Baseline and Truth/Deception for all speakers

3.3 Vowel Formants

F1

Figure 3 illustrates the mean F1 values for each measured vowel category from each individual speaker. The x-axis represents the F1 taken from the Baseline condition and the difference between Baseline and Truth/Deception is shown along the y-axis. The majority of tokens lay on or slightly above the origin on the y-axis, which indicated that the F1 values from Truth/Deception were similar to those from the Baseline speech. There was some variability across tokens, demonstrated by the moderate spread of data points. As suggested by the almost horizontal trend lines, no significant correlations existed between F1 in the Baseline condition and the change in F1 in the Truth (r = 0.03926, df = 38, p = 0.8099) or Deception conditions (r = 0.00948, df = 38, p = 0.9537).

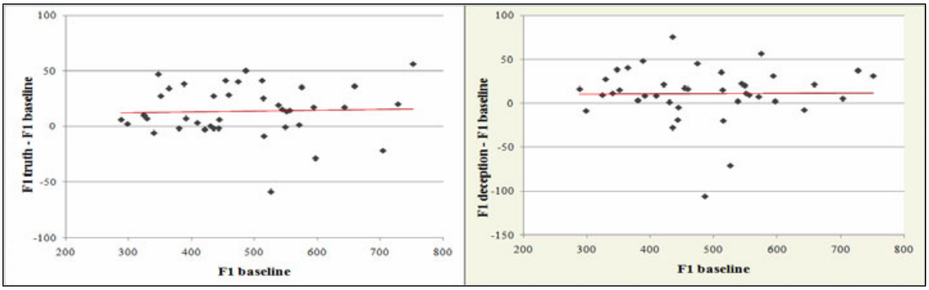


Fig. 3. Scatter-plot of F1 measures, showing value in the Baseline condition (x-axis) against shift observed in the Truth/Deception condition (y-axis)

F2

Figure 4 reflects the observed directional inconsistencies for F2. Some values were slightly increasing, some were decreasing and others were not changing. Overall the change was not considerable for any of the vowel categories and T-tests did not illustrate any significant differences at the 5% level.

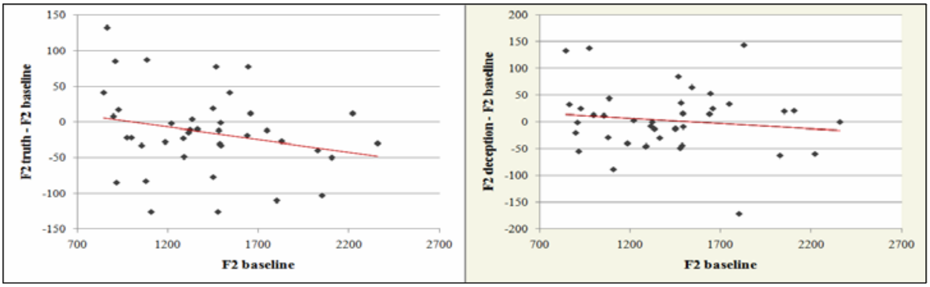


Fig. 4. Scatter-plot of F2 measures, showing value in the Baseline condition (x-axis) against shift observed in the Truth/Deception condition (y-axis)

In particular for the Truth condition, there was a substantial amount of variation between tokens in the size of the difference as well as the direction. The same vowel category might be increasing, decreasing or not changing between different speakers. There was less variation in the Deception condition and most of the tokens were grouped around the origin. The nearly horizontal trend-line reinforces the observation that no real pattern can be detected. The correlation between Baseline vowel measurements and the effect of Truth (-0.24396, df = 38, p = 0.1292) and Deception (r = -0.12698, df = 38, p = 0.4349) was weak and not statistically significant at any conventional significance level.

F3

The F3 values of Baseline and Truth/Deception seemed to correspond very closely to each other and there did not appear to be a difference for any of the vowel categories. If changes did occur then it tended to be a decrease in Truth and Deception compared to Baseline.

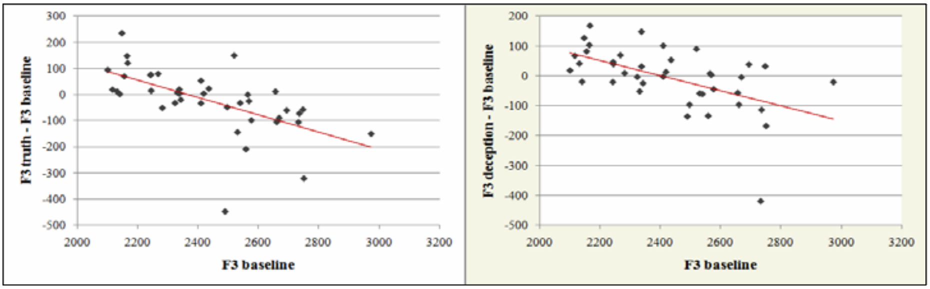


Fig. 5. Scatter-plot of F3 measures, showing value in the Baseline condition (x-axis) against shift observed in the Truth/Deception condition (y-axis)

There was a tendency that high F3 values in the Baseline condition were more likely to be subject to a decrease in Truth/Deception than low F3 values. The overall correlation between F3 in Baseline and F3 decrease in Truth (r = -0.59525, df = 38, p < .001) and Deception (r = -0.54605, df = 38, p < .001) was statistically significant.

4 Discussion

The results suggest that truth-tellers and liars cannot be differentiated based on the speech signal measures analysed in this study. Not only was there a lack of significant changes for the majority of parameters investigated but also, if change was present it failed to reveal consistencies within and between the speakers.

F_0 mean varied little across conditions. The reduced F_0 SD values in T and D suggest that speakers were less variable and perhaps spoke with a more monotone voice in these conditions. This could be further investigated by auditory analysis.

With regard to overall intensity changes, the findings also did not offer grounds for a reliable distinction between those telling the truth or lying. If speakers showed a change in mean energy then it was uniform in terms of direction and size across Truth and Deception.

The majority of F1 and F2 differences between conditions were not statistically significant. For F2 in particular, there was a considerable amount of variation across conditions with values increasing, decreasing or not changing. The F3 results point towards a negative correlation between Truth/Deception and Baseline speech. Of note again is the parallelism between Truth and Deception in that both show a significant negative correlation compared to the Baseline. F3 is linked to voice quality and vocal profile analysis of the speakers would cast more insights.

The remarkable amount of inter- and intra-speaker variability underlines the fact that deceptive behaviour is individualised, very multifaceted and far from being clear cut and straightforward. Despite their non-significance the findings are of interest since they point to some potential limitations when trying speech analysis for deception detection purposes only.

It may be argued that the lack of significant findings is a product of the experimental arrangement as a laboratory induced deception which does not adequately represent deception as it might occur in real life. This is a methodological limitation which, due to ethical considerations, cannot be overcome in the majority of studies on deception. In order to maintain the impact of the scientific validity of this study, it can be said that post-interview rating scales confirmed that all the participants were highly motivated to succeed in the deceptive act (score of 6 or higher on a 7- point Likert scale). It should also be stated that research into a relatively unexplored area, such as speech and deception, needs to start off with fully controlled experiments where variables can be controlled more strictly. Clean, high quality recordings must provide the starting point for the acoustic and phonetic analysis. If differences between truth and deception are found in these ideal conditions, research can then move on to investigating less controlled data in the field.

One of the assets of the present research design concerns the separation of stress and deception in that the latter was not inferred from the former. The polygraph and most of the VSA technologies are based on the assumption that liars will show more emotional arousal i.e. will experience more stress than truth-tellers [2]. However, such a direct linkage cannot be presupposed. Certainly, there will be liars who do manifest the stereotypical image of nervousness and stress. At the same time, however, truth-tellers may also exhibit anxiety and tension, especially if in fear of not being trusted. And on the contrary, liars might not conform to the stereotypical image described above but rather display a composed and calm countenance. As the following quote illustrates:

> *'Anyone driven by the necessity of adjudging credibility who has listened over a number of years to sworn testimony, knows that as much truth has been uttered by shifty-eyed, perspiring, lip-licking, nail-biting, guilty- looking, ill-at-ease fidgety witnesses as have lies issued from calm, collected, imperturbable, urbane, straight-in-the-eye perjurers.'(Jones, E.A. in [2, p.102])*

Harnsberger et al. [5] for example only included participants in their analysis who showed a significant increase in stress levels during deception. Given that the goal of their research was to test the validity of VSA technology this may be a justified methodological choice. However, as the aim of the present study was to attain a more comprehensive knowledge of the fundamental relationship between deception and speech it was essential to disassociate deception and stress.

Further acoustic and phonetic analysis is under way to expand the analysis beyond measures of f_0, intensity, and formant measurements to include measurement of diphthong trajectories, consonant articulation, jitter, shimmer and spectral tilting. In addition, laryngograph recordings have been made with 10 speakers which will provide opportunity to analyse the glottal wave signal and this in turn will contribute further to our knowledge of truth/deception specific speech characteristics. Furthermore, the hypothesis that increasing cognitive load during interview situations has the potential of magnifying the differences between truth-tellers and liars in the speech domain will also be evaluated.

5 Conclusion

This paper summarised an exploratory investigation into the relationship between acoustic parameters of speech and truth/deception. So far the analysed data does not suggest that a reliable and consistent correlation exists. As well as providing a basis for future research programs the present study should encourage researchers and practitioners to evaluate critically what is and what is not possible, using auditory and machine based analyses, with respect to detecting deception from speech.

Acknowledgement. This work was made possible through EPSRC Grant number: EP/H02302X/1. Thanks to Philip Harrison for providing the Praat script and Francis Newton for his time and assistance during the data collection process.

References

1. Jessen, M.: Einfluss von Stress auf Sprache und Stimme. Unter besonderer Beruecksichtigung polizeidienstlicher Anforderungen. Schulz- Kirchner Verlag GmbH, Idstein (2006)
2. Lykken, D.: A Tremor in the Blood: Uses and Abuses of the Lie Detector. Perseus Publishing, Reading (1998)
3. Damphousse, K.R., Pointon, L., Upchurch, D., Moore, R.K.: Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting. Report submitted to the U.S. Department of Justice (2007)
4. Eriksson, A., Lacerda, F.: Charlantry in forensic speech science: A problem to be taken seriously. International Journal of Speech, Language and the Law 14(2), 169–193 (2007)
5. Harnsberger, J.D., Hollien, H., Martin, C.A., Hollien, K.A.: Stress and Deception in Speech: Evaluating Layered Voice Analysis. Journal of Forensic Science 54(3), 642–650 (2009)
6. Benus, S., Enos, F., Hirschberg, J., Shriberg, E.: Pauses in deceptive Speech. In: Proceedings ISCA 3rd International Conference on Speech Prosody, Dresden, Germany (2006)
7. Feeley, T.H., deTurck, M.A.: The behavioural correlates of sanctioned and unsanctioned deceptive communication. Journal of Nonverbal Behavior 22(3), 189–204 (1998)
8. Stroemwall, L.A., Hartwig, M., Granhag, P.A.: To act truthfully: Nonverbal behaviour and strategies during a police interrogation. Psychology, Crime and Law 12(2), 207–219 (2006)
9. Anolli, L., Ciceri, R.: The Voice of deception: Vocal strategies of naïve and able liars. Journal of Nonverbal Behavior 21(4), 259–284 (1997)

10. Rockwell, P., Buller, D.B., Burgoon, J.K.: The voice of deceit: Refining and expanding vocal cues to deception. Communication Research Reports 14(4), 451–459 (1997)
11. Enos, F.: Detecting Deception in Speech. PhD Thesis submitted to Columbia University (2009)
12. Porter, S.B.: The Language of Deceit: Are there reliable verbal cues to deception in the interrogation context? Master's thesis submitted to The University of British Columbia (1994)
13. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (Computer program). Version 5.2.12, http://www.praat.org/ (retrieved January 28, 2011)
14. Wells, J.C.: Accents of English I: An Introduction. Cambridge University Press, Cambridge (1982)

Kansei Evaluation of HDR Color Images with Different Tone Curves and Sizes – Foundational Investigation of Difference between Japanese and Chinese

Tomoharu Ishikawa[1], Yunge Guan[1], Yi-Chun Chen[1], Hisashi Oguro[1, 2],
Masao Kasuga[1], and Miyoshi Ayama[1]

[1] Graduate School of Engineering, Utsunomiya University,
7-1-2 Yoto Utsunomiya Tochigi, 321-8585, Japan
[2] TOPPAN PRINTING Co., Ltd., 1-3-3 Suido, Bunkyo, Tokyo
112-8531, Japan
ishikawa@is.utsunomiya-u.ac.jp,
{mt106623,dt097173}@cc.utsunomiya-u.ac.jp,
Hisashi.Oguro@toppan.co.jp,
{kasuga,miyoshi}@is.utsunomiya-u.ac.jp

Abstract. High dynamic range (HDR) color images are evaluated for Kansei impression by two groups of observers: Japanese and Chinese. Twenty HDR images were created by converting each of five HDR images with different tone curve properties into four screen sizes. As a result, the subjective rating value for the psychophysical properties of images, such as “Light” and “Dark,” increased or decreased monotonically with the average brightness *L**, but not with image size. On the other hand, the rating value for some Kansei evaluations, including “Natural” and “Clear,” followed the same patterns. Next, we applied factor analysis to the results, having divided the data into Japanese and Chinese. The analysis result indicated that two and three factors were extracted from the rating value evaluated by Chinese and Japanese participants, respectively. These results suggest that Japanese observers evaluated HDR images in more detail than Chinese ones did.

Keywords: Kansei Evaluation, High Dynamic Range, Japanese and Chinese, Tone Curve, Screen Size.

1 Introduction

Currently, various types of image contents are being delivered all over the world though the Internet. The quality of these images is judged on the basis of observers' experience and knowledge, and, for example, the quality of image gradation and resolution and screen size [1]–[3]. In many cases, observers' impressions do not agree, even when they receive the same content. This is a serious problem for creating Web content. In previous studies on this problem, the relationship between the image qualities, including lightness contrast and image size, and the effect of value judgment upon Kansei impression have been investigated [4], [5]. The results showed that

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 38–44, 2011.

adjectives were divided into three groups. The first and second groups, which were related to psychophysical properties, were strongly affected by lightness contrast and image size, respectively. On the other hand, the third group, which was related to the Kansei impression, was affected by both lightness contrast and image size. This study aims to examine whether comparable results can be obtained when an image of a different type is evaluated, and what difference appears in evaluations by observers of different national origins. Specifically, high dynamic range (HDR) images [6], which have a relatively fine gradation of the image, are evaluated for Kansei impression by two groups of observers—Japanese and Chinese—who are assumed to have different value judgments. The results show that the subjective rating value for psychophysical properties of the image, such as “Light,” “Dark,” “Deep color,” and “Pale color,” increased or decreased monotonically with average brightness *L**, but not with image size. On the other hand, the rating value for some Kansei evaluations, including “Natural,” “Unnatural,” “Clear,” and “Vague,” followed these same patterns. Next, we applied factor analysis to the results, having divided the data into Japanese and Chinese. The result indicated that two and three factors were extracted from the rating value evaluated by Chinese and Japanese participants, respectively. This indicates that Japanese observers evaluated HDR images in more detail than Chinese ones did.

2 Experiment

2.1 Experimental Stimuli

Nightscape images were taken by a digital camera (Nikon D50) whose exposure values were set into ten steps. An HDR image was created by the synthesis of these photos using Photomatix Pro 3.0, and then, five HDR images were generated with Photoshop CS4 by changing the tone curves. Two kinds of image contents are shown in “Amusement Park” (Fig. 1) and “Shopping Mall” (Fig. 2). Twenty HDR images, made by converting each of the five HDR images into four screen sizes (7, 14, 29, and 57in), were employed as the experimental stimuli.

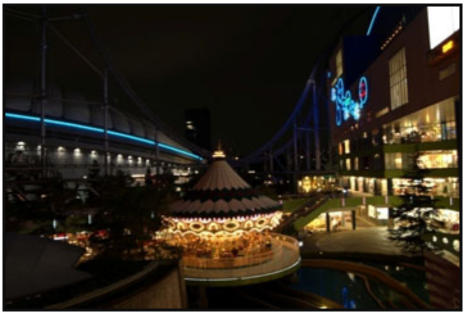


Fig. 1. Amusement Park

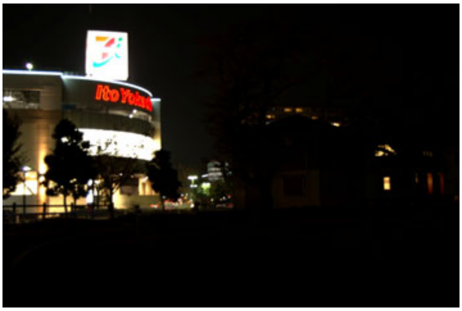


Fig. 2. Shopping Mall

2.2 Experimental Conditions and Procedure

Observers were 24 participants with normal color vision: 12 Chinese and 12 Japanese. An experimental booth was constructed by covering walls with gray curtains. Ambient light was provided with a fluorescent light fixture in the ceiling of the room. Horizontal and vertical illuminances near the center of the display were about 285 and 350 lx, respectively. A participant entered the booth and was allowed 3 min to adapt to the visual environment before being instructed to evaluate the experimental stimulus by choosing his/her evaluation using a seven-step scale, from 0 to 6, for each adjective listed on the answer sheet. The experimental stimulus was presented on a 65-in display (SHARP AQUOS LC-65RX1W), and the participant was given as much time as desired to carry out the evaluation. In order to avoid comparison between successive images, a homogeneous gray plane (N5) was presented between each pair of HDR images for 5 s. Two visual distances were used: 160 and 320 cm. The unipolar scale method was employed using 22 adjectives. These adjectives (Table 1) were provided in the native language of the participant. The appearance of the experiment is shown in Fig. 3.

Table 1. Twenty-two adjectives

Psychophysical properties	Strong contrast	Weak contrast	Light	Dark
	Deep color	Pale color		
Kansei evaluations	Clear	Vague	Showy	Plain
	Natural	Unnatural	Clean	Dirty
	Like	Hate	Easy to view	Difficult to view
	Stereoscopic	Flat	Impressive	Ordinary
	*Dynamic	*Static	*Amusing	*Uninteresting

*These adjectives were added to the evaluation of “Amusement Park”.

Fig. 3. Appearance of the experiment

3 Result and Analysis

The average of the subjective rating values for each image, adjective, observer group, image size, and visual distance was calculated. In several cases, the degree of change in the subjective rating value with an increase in the average brightness *L** for Image-I was larger than that for Image-II, and the rating value for Chinese participants was larger than that for Japanese ones. In other words, the result indicated that the differences in subjective rating value for the observers and image contents are influenced by the *L**value. Fig.4 shows the subjective rating values of the assessment word for “Light” of Japanese participants for a visual distance of 160cm using Image-I. As shown in the figure, the subjective rating values increase with an increase in average brightness *L** for all image sizes. On the other hand, the rating values for “Dark” decrease with the increase in the *L**value for all image sizes. These results are nearly symmetrical and both show monotonous change with changing *L**value. The results for different image sizes show a similar tendency, indicating that image size has no particular effect on the assessment of “Light” and “Dark.” Similar results are obtained for the assessment words for “Deep color” and “Pale color.” These words express the psychophysical properties of the image. Therefore, the evaluation of psychophysical properties of an image simply changes with an increase in the *L**value.

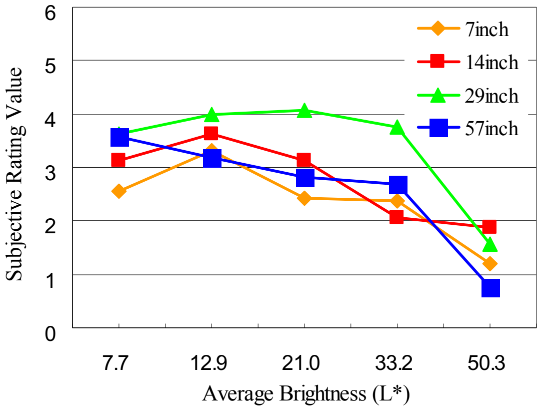


Fig. 4. Subjective rating values for “Light” with a visual distance of 160 cm for Image-I, rated by Japanese participants

The subjective rating value for the assessment word for “Natural” indicated in Fig.5(a) shows a maximum at the *L**value (12.9 or 21.0) for the 7, 14, and 29in screens by Japanese participants. Results similar to Fig.5(a) are observed in other Kansei evaluations, e.g., for “Clear.” Although Fig.5(b) shows the result under the same conditions but with Chinese participants, the *L**value (21.0 or 33.2) at which the maximum is attained is higher than that for Japanese ones. Results similar to those in Fig.5(b) are observed in Kansei evaluations, e.g., for “Like.” Therefore, Kansei evaluations show different trends than psychophysical evaluations do. Moreover, the rating value of “Easy to view” shows a size dependence in both observer groups; however, the degree of dependence is significantly greater among Chinese rather than among Japanese participants. These results indicate that the evaluation of HDR image approximately corresponds to the results of previous research.

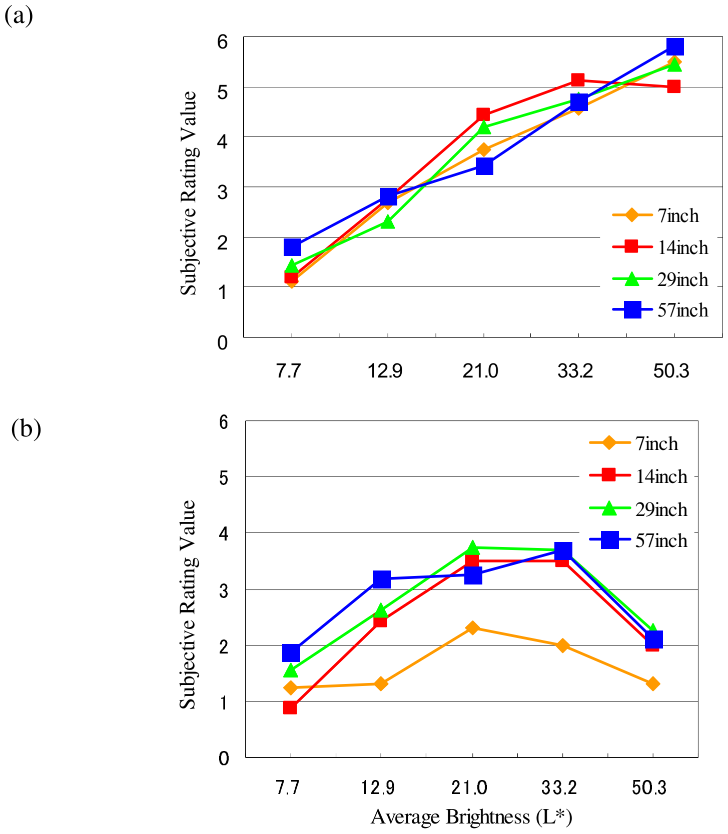


Fig. 5. (a) Subjective rating values for “Natural” with a visual distance of 160 cm in the case of Image-I as rated by Japanese participants. (b) Corresponding rating values of Chinese participants.

In order to extract the factors that contribute to the Kansei evaluation, we applied factor analysis to the 40 results (20 images × 2 distances; in each of them, the average

Table 2. Results of factor analysis for Image-II

	Japanese			**Chinese**	
Adjectives	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 1**	**Factor 2**
Easy to view	**0.87**	0.07	0.35	**-0.96**	0.04
Like	**0.86**	0.23	-0.18	**-0.97**	0.14
Clear	**0.85**	-0.33	-0.15	**-0.95**	0.19
Impressive	**0.81**	-0.43	0.09	**-0.96**	0.15
Clean	**0.80**	-0.08	-0.31	**-0.95**	0.22
Stereoscopic	**0.74**	-0.25	0.52	**-0.96**	0.09
Flat	**-0.75**	0.16	-0.37	**0.91**	-0.15
Hate	**-0.81**	-0.38	0.15	**0.96**	-0.17
Difficult to view	**-0.90**	-0.14	-0.20	**0.97**	-0.08
Ordinary	0.27	**0.86**	-0.11	**0.94**	-0.19
Plain	-0.27	**0.80**	-0.43	**0.87**	-0.33
Natural	0.59	**0.74**	-0.20	**-0.94**	0.00
Dark	-0.30	**0.69**	-0.59	**0.87**	-0.40
Light	0.34	**-0.73**	0.53	**-0.86**	0.42
Unnatural	-0.49	**-0.74**	0.34	**0.93**	-0.03
Showy	0.30	**-0.83**	0.40	**-0.95**	0.22
Vague	-0.07	-0.06	**0.83**	**0.97**	-0.12
Dirty	-0.42	-0.40	**0.63**	**0.95**	-0.20
Weak contrast	0.11	-0.41	**0.82**	0.09	**0.51**
Pale color	0.13	-0.60	**0.73**	-0.22	**0.79**
Strong contrast	0.01	0.19	**-0.75**	0.10	**-0.77**
Deep color	-0.04	0.48	**-0.82**	0.49	**-0.67**

scores of 12 participants were used), having divided the data into Japanese and Chinese. After the factor extraction, varimax rotation was used following the main factor method. The results for the Japanese participants indicated that three factors, with an eigenvalue larger than 1, were extracted, and their cumulative contribution rates were larger than 80%; in contrast, two factors were obtained in the case of Chinese participants. The results of factor analysis for Image-II are listed in Table 2. In the case of Japanese participants, vision perception characteristics, including "Stereoscopic" and "Easy to view," showed the largest values for factor loading in the first factor, and these were extracted as the "Vision perception factor." In the second factor, impressive quality, including "Natural" and "Showy," had the largest values for factor loading; this factor was named the "Impressive quality factor." The third factor was called the “Contrast factor,” as its elements were “Deep color” and “Strong contrast.” On the other hand, the first factor for Chinese participants was almost identical to the first and second factors for Japanese ones, and its second factor was almost identical to the third factor for the Japanese group. These results suggest that Japanese participants evaluate HDR images in more detail than Chinese ones do.

4 Conclusion

In this study, to clarify the difference in Kansei evaluation by observers of different national origins—Japanese and Chinese—HDR images were evaluated for Kansei impression. The results showed that the subjective rating value for psychophysical properties of the image, such as “Light,” “Dark,” “Deep color,” and “Pale color,” increased or decreased monotonically with average brightness L^*, but not with image size. On the other hand, the rating value for some Kansei evaluations, including “Natural,” “Unnatural,” “Clear,” and “Vague,” followed the same pattern. These results indicate that the evaluation of HDR image approximately corresponds to the results of previous research. Moreover, we applied factor analysis to the results, having divided the data into Japanese and Chinese. The result indicated that two and three factors were extracted from the rating values generated by Chinese and Japanese participants, respectively. These results suggest that Japanese participants evaluated HDR images in more detail than Chinese ones did.

Acknowledgment. This research was supported by Eminent Research Selected at Utsunomiya University.

References

1. Duchnicky, R.L., Kolers, P.A.: Readability of Text Scrolled on Visual Display Terminals as a Function of Window Size. Human Factors 25, 683–692 (1983)
2. Chae, M., Kim, J.: Do Size and Structure Matter to Mobile Users? An Empirical Study of the Effects of Screen size, Information Structure, and Task Complexity on User Activities with Standard Web Phones. Behaviour and Information Technology 23(3), 165–181 (2004)
3. Hatada, T., Sakata, H., Kusaka, H.: Induced Effect of Direction Sensation and Display Size: Basic Study of Realistic Feeling with Wide Screen Display. The Journal of the Institute of Television Engineers of Japan 33(5), 407–413 (1979) (in Japanese)
4. Ishikawa, T., Shirakawa, T., Oguro, H., Guo, S., Eda, T., Sato, M., Kasuga, M., Ayama, M.: Color Modulations Appropriate for Different Image Size Based on Impression Assessment. In: 11th Congress of the International Colour Association AIC 2009, Sydney (2009)
5. Chen, Y.-C., Ishikawa, T., Shirakawa, T., Eda, T., Oguro, H., Guo, S., Sato, M., Kasuga, M., Ayama, M.: Effects of Lightness Contrast and Image Size on KANSEI Evaluation of Photographic Pictures. Kansei Engineering and Emotion Research KEER 2010, Paris (2010)
6. Horiuchi, T., Fu, Y.Q., Tominaga, S.: Perceptual and Colorimetric Evaluations of HDR Rendering with/without Real-world Scenes. 11th Congress of the International Colour Association AIC 2009, Sydney (2009)

Personalized Emotional Prediction Method for Real-Life Objects Based on Collaborative Filtering

Hyeong-Joon Kwon, Hyeong-Oh Kwon, and Kwang-Seok Hong

School of Information and Communication Enginnering, Sungkyunkwan University,
300, Chunchun-dong, Jangan-gu, Suwon, Kyungki-do, 440-746, South Korea
{katsyuki,oya200}@skku.edu, kshong@skku.ac.kr

Abstract. In this paper, we propose a personalized emotional prediction method using the user's explicit emotion. The proposed method predicts the user's emotion based on Thayer's 2-dimensional emotion model, which consists of arousal and valence. We construct a user-object dataset using a self-assessment manikin about IAPS photographs, and predict the target user's arousal and valence by collaborative filtering. To evaluate performance of the proposed method, we divide the user-object dataset into a test set and training set, and then observe the difference between real emotion and predicted emotion in the 2-dimensional emotion model. As a result, we confirm that the proposed method is effective for predicting the user's emotion.

Keywords: Emotional Prediction, IAPS, Self-assessment Manikin.

1 Introduction

One of the main study areas of the HCI field is building a computer with human communication abilities. This has been preceded by a one-way interface to mutual interaction based on physical communication between a computer and human. As a result, computer vision and speech recognition (which provides technology with the human ability to recognize objects and humans using eyes and ears) took a major step forward.

Recent studies reproduce human sense organs such as eyes and ears by computer using pattern recognition and machine learning technology. In addition, more studies on human emotion recognition (called emotion computing) is in progress. The method based on sight is processed with a face visual to determine different features of emotion from one's facial expressions [10], and is recognized by computer. Javier Movellan has announced that a computer can determine great and small changes from every part of a human face to recognize emotion such as anger, sadness, displeasure, pleasure, and more [1]. The method based on hearing uses different human voice features from emotional changes. Speech signals differ for each individual for distinct articulator and speaking habits, so usual voice signals and other voice signals with features are detected to recognize emotion [8][9].

However, these approaching methods involve several limitations. First, visual and voice methods are useful to recognize outward appearance constituents. But emotion is the constituent not exposed by outward appearance, so there is always a limitation

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 45–52, 2011.

from using visual and voice methods. Second, facial expressions and human voices differ for all individuals on each emotion. Humans may express the exact facial expression and voice to represent different emotions, and emotions toward the same object may vary. This means that existing methods using visual and vocal features cannot find the sensibility of an individual, and have limitations in emotion recognition. Human emotions differ depending on individual sensibilities, so the computer must reflect individual sensibilities for more accurate emotion recognition. For example, an object may cause one person to feel sad, while another encounter with the same object may not cause sadness; this is because of the difference in sensibilities due to individual experience. Therefore, the study of emotion prediction or recognition methods based on the sensibilities of individuals is necessary.

In this paper, we propose a personalized emotional prediction method for real-life objects, IAPS photographs [2], based on and SAM [3] and collaborative filtering [5] which uses individual tendencies. Proposed method uses emotion data of users with similar sensibilities compared to a target user on a certain object to predict the target user's emotion. This emotion recognition method uses the convergence of an emotion induction dataset and soft computed feelings of computer science built as emotion based on psychological basis. Performance evaluation criteria are 1) Lineal distance of subjective emotion of actual user and emotion predicted using proposed method, and 2) The mean error between subjective emotion evaluation score rated by user and emotion evaluation predicted with proposed method.

This paper is organized as follows. In section 2, we describe IAPS and collaborative filtering to design the proposed method. In section 3, we propose a personalized emotional prediction method using memory-based collaborative filtering with explicit SAM rating on IAPS photographs. In section 4, we show the prediction accuracy of SAM assessment and the overall performance of emotional prediction. In section 5, we summarize the results of this study and suggest topics for future work.

2 Related Works

In this section, we describe the IAPS photographs dataset that promotes human emotion, an effective method for evaluating explicit emotion (SAM), and memory-based collaborative filtering (MBCF), which is a collective intelligence algorithm in the realm of artificial intelligence. We design and implement a personalized emotional prediction method using IAPS photographs, CF, and SAM in section 3.

2.1 IAPS and SAM

The Interactive Affection Picture System (IAPS) includes 1,200 photographs that stimulate human emotions [2]. Existing studies measure emotion toward IAPS photographs via SAM, which consists of pressure, arousal, and dominance. SAM contains various merits. First, SAM doesn't contain a language. Thus, pure emotion is measured without nationality. Second, even illiterate users can respond emotionally because SAM is designed as a simple picture. Third, because same pictures about each element are used, it is possible between different cultural areas. Fig. 1 shows our SAM format, which contains valence and activation [3].

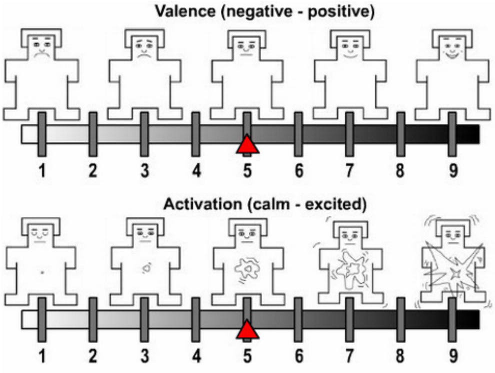


Fig. 1. Self-assessment Manikin

2.2 CF: Collaborative Filtering

The CF is a famous collective intelligence algorithm. It discovers similar points among people and predicts a preference rating for an unfamiliar item between similar people. Modern recommender systems are based on CF. The CF is divided into memory-based CF and the model-based approach [5].

1. One of the model-based CF approaches uses a clustering algorithm. This method divides all users in a user-item rating matrix into several groups. Clustering is a technique widely used for the statistical analysis of data, and is an important unsupervised learning method. To identify interesting data distributions and patterns, clustering techniques classify physical or abstract objects into classes such that the objects in each class share some common attribute. Depending on the characteristics of the distinct clusters, companies can make independent decisions about each cluster. The model-based CF is robust to cold-start problems, but it is a highly vulnerable real-time prediction because of the training process.
2. MBCF calculates the similarity between users or items based on the user-item rating matrix [4]. To predict user m for item n, the user-based approach calculates the similarity of co-ratings from two users, and arrays all other users in order of similarity to user m. The item-based approach then calculates the similarity between two items using their co-ratings, and arrays all other items in order of similarity. The MBCF is robust to real-time recommendation, but has an increased calculating cost. We predict arousal and valence ratings based on this method in this paper.

3 Proposed Emotional Prediction Method

Fig. 2 shows the structure of a user-object dataset and Thayer's 2-dimensional emotion model. On the left, n indicates the IAPS photograph number, m indicates a user, and r means rating. The proposed method contains two matrices like Fig. 2: arousal and valence. We can know a user's emotion toward a photograph based on the arousal and valence coordinates. Thayer's 2-dimensional emotion model shows the user's emotion using a pair of arousal and valence coordinates about a photograph. Dominance is not used in Thayer's 2-dimensional emotion model.

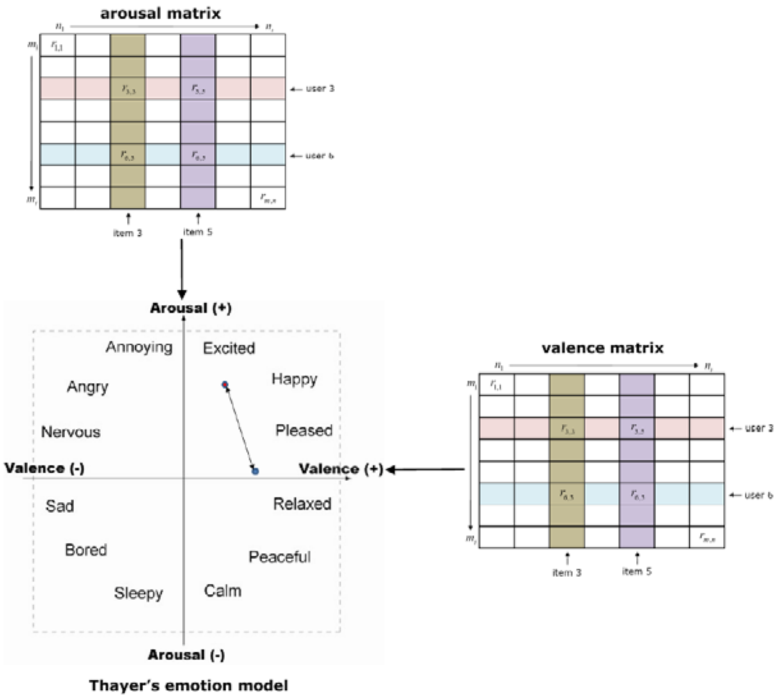


Fig. 2. A structure of a user-object dataset and Thayer's 2-dimensional emotion model

A generalized step of the proposed method is described as follows:

1. It collects arousal and valence ratings about various real-life objects from a great number of users, and then constructs a user-object dataset, which consists of an arousal matrix and valence matrix. We use IAPS photographs on behalf of real-life objects. Also, we developed software to collect SAM ratings about IAPS photographs.
2. It chooses a target user who is somebody to predict emotion. The target user must have been evaluates objects in user-object dataset. This is a necessary condition to use collaborative filtering.
3. It shows the target object to the target user. Then, it searches *top-k* similar users which is rated target object. When it searches similar users, it considers linear similarity algorithms such as PCC, COS, RMS, and so on. The vector space model-based method includes Cosine Similarity (COS), which is frequently used in information retrieval. The COS assumes that the rating of each user is a point in vector space, and then evaluates the cosine angle between the two points. It considers the common rating vectors $X = \{x_1, x_2, x_3, \cdots, x_n\}$ and $Y = \{y_1, y_2, y_3, \cdots, y_n\}$ of users X and Y. It is represented by a dot-product and magnitude. COS has frequently been used for performance comparisons in the CF area. The cosine angle between vectors X and Y is given by:

$$\cos(X,Y) = \frac{X \bullet Y}{\|X\|\|Y\|} = \frac{x_1 y_1 + x_2 y_2 \hbar \ x_n y_n}{\sqrt{x_1{}^2 + y_1{}^2}\sqrt{x_2{}^2 + y_2{}^2}\hbar \ \sqrt{x_n{}^2 + y_n{}^2}} \tag{1}$$

One correlation-based method is the Pearson dot-product Correlation Coefficient (PCC). This method is normally used to evaluate the association intensity between two variables. PCC is given by:

$$\gamma(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}} \tag{2}$$

Row moment-based similarity (RMS) is divided into three steps based on the rating matrix [6]. The first step is target user profiling, which evaluates the similarity between the target user and everybody else. Assume that the target user is u_3. The similarity between u_3 and u_5 is evaluated by co-ratings composed of the user's ratings for the same contents. The $r_{3,4}$ and $r_{5,4}$ values are one of the co-ratings between u_3 and u_5. The COS and PCC are common similarity algorithms in the MBCF study area. This similarity is faster and simpler than well-known existing algorithms.

$$\begin{aligned} RMS(u_i, u_j) &= 1 - \frac{1}{r^k}\frac{1}{n}\sum_{v=1}^{n}(|u_{i,v} - u_{j,v}|)^k \\ &= 1 - \frac{1}{r^k}\sum_{z=1}^{m}\Pr(D = d_z)\cdot d_z{}^k \end{aligned} \tag{3}$$

4. It predicts valence and arousal ratings based on a user-object dataset. It then maps arousal and valence coordinates to Thayer's 2-dimensional emotion model. This shows the target user's emotion toward the target object. After calculating the similarity for all pairs of objects, the MBCF selects *top-k* similar users. Accordingly, the MBCF predicts the rating of the target user for the target item with similarities for weights. (4) shows the prediction method, where *u* indicates the target user, *i* indicates the target item, *j* indicates other users, and *w* is similarity between *u* and *j* [4].

$$p(u,i) = \overline{r}_u + \frac{\sum_{j=1}^{n} w_{j,v}(r_{j,i} - \overline{r}_j)}{\sum_{j=1}^{n} w_{u,j}} \tag{4}$$

5. It measures prediction accuracy. Mean absolute error is used between predicted valence and arousal ratings and real valence and arousal ratings. In (5), p indicates prediction rating and *q* indicates real rating, while *n* represents the total prediction counts.

$$MAE = \frac{\sum_{i=1}^{n}|p_i - q_i|}{n} \tag{5}$$

6. It maps a point using prediction arousal and valence ratings on Thayer's 2-dimensional emotion model, and then maps a point using real arousal and valence ratings on the same emotion model. Next, it measures the Euclidean distance between the real point and prediction point. This distance is the major performance of the proposed method. The two points in Fig. 2 represent this.

4 Experimental Results

We have collected subjective emotions of individual users using SAM over IAPS of an often used emotion induction dataset from 69 users to build individual emotion sensibility data. IAPS consists of 1,200 photographs of commonly seen objects from human life. Extremely suggestive or cruel photographs were excluded; 1,073 photographs were composed randomly into 8:2 ratios (80 % of photographs as training data and 20 % as test data).

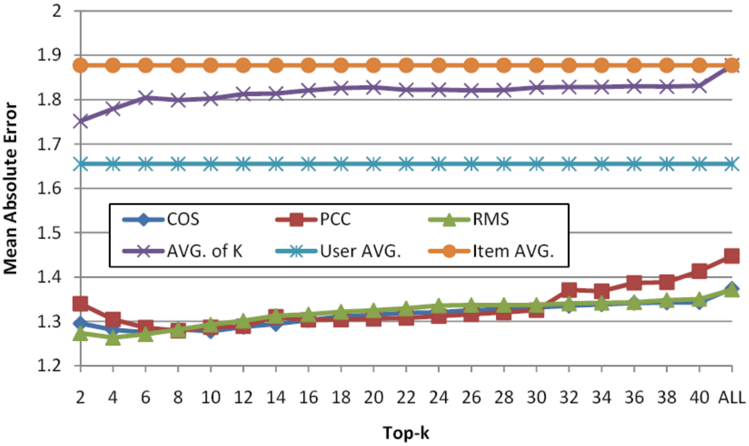


Fig. 3. Arousal prediction result

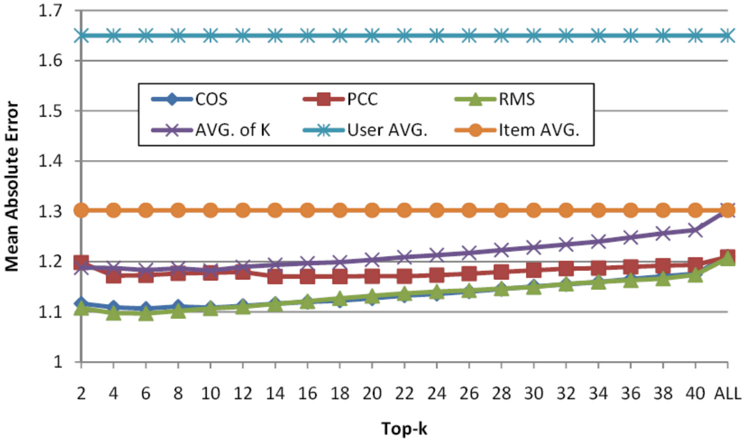


Fig. 4. Valence prediction result

We then used PCC, COS, RMS, and *top-k* user's rating average to predict 20 % test data ratings. The prediction results are shown in Fig. 3 and Fig. 4. The MAE (absolute average of the difference between real rating and prediction rating) curve for MBCF descends with increasing neighbors and rises at the optimal point. The experimental result shows a remarkable result in arousal and valence datasets. The Proposed method showed greater prediction accuracy than *top-k* average.

Table 1. Emotion prediction result based on Thayer's 2-dimensional model

Similarity method	Optimal *top-k*	Error distance
COS	6	**2.3477**
PCC	14	2.6607
RMS	6	2.3624
AVG. of k	10	2.3489
User AVG.	-	2.5998
Item AVG.	-	2.8610

Next experiment, in Table 1, is to observe the error distance of a 2-dimensional emotion model. This experiment predicts arousal and valence ratings about target objects, and marks real values and predicted values on Thayer's 2-dimensional emotion model. We then measured the lineal absolute distance error between the two points. Euclidean distance was used as a distance measurement. Table 1 shows the result and experimental condition which contains optimal *top-k* value per similarity method.

5 Conclusion

We proposed an emotion prediction method using MBCF. The proposed method predicts arousal and valence ratings based on a user-object dataset and discovers user emotions based on Thayer's 2-dimensional emotion model. From the experimental results, the proposed method showed greater prediction accuracy than a simple average. In the future, we will study a multimodal emotion recognition method using face or speech. This approach will be an optimal recognition method using external and internal human elements.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0021411), and by MKE, Korea under ITRC NIPA-2010-(C1090-1021-0008) (NTIS-2010-(1415109527)).

References

1. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Movellan: Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In: The 2003 IEEE Conference on International Conference on Computer Vision and Pattern Recognition Workshop, pp. 1–6. IEEE Press, Nwe York (2003)

2. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International Affective Picture System (IAPS): Technical Manual and Affective Ratings. NIMH Center for the Study of Emotion and Attention (1997)
3. Bradley, M.M., Lang, P.J.: Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25(1), 49–59 (1994)
4. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An Algorithmic Framework for Performing Collaborative Filtering. In: The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999), pp. 230–237. ACM Press, New York (1999)
5. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)
6. Kwon, H.-J., Hong, K.-S.: Moment Similarity of Random Variables to Solve Cold-start Problems in Collaborative Filtering. In: Third International Symposium on Intelligent Information Technology Application, pp. 584–587. IEEE Press, New York (2009)
7. Ahn, H.J.: A New Similarity Measure for Collaborative Filtering to Alleviate the New User Cold-start Problem. Information Science 178(1), 37–51 (2008)
8. New, T.L., Foo, S.W., De Silva, L.C.: Speech Emotion Recognition using Hidden Markov Models. Speech Communication 41(4), 603–623 (2003)
9. Roh, Y.-W., Kim, D.-J., Lee, W.-S., Hong, K.-S.: Novel Acoustic Features for Speech Emotion Recognition. Science in China Series E: Technological Sciences 52(7), 1838–1848 (2009)
10. De Silva, L.C., Miyasato, T., Nakatsu, R.: Facial Emotion Recognition Using Multi-modal Information. In: The International Conference on Information, Communications and Signal Processing ICICS 1997, pp. 397–401. IEEE Press, Los Alamitos (1997)

A Study of Vision Ergonomic of LED Display Signs on Different Environment Illuminance

Jeih-Jang Liou, Li-Lun Huang, Chih-Fu Wu, Chih-Lung Yeh, and Yung-Hsiang Chen

The Graguate Institute of Design Science, Tatung University

Abstract. The LED (light emitting diode, also referred to as LED) have already been used widely. However, despite the high visibility of LED with high brightness performance, it also leads to a glare problem, which generates a direct security issue in applying to traffics. Therefore, this research aimed to study how to make the LED display sign be more legible under high illuminative environments and to avoid the observers feeling dazzling glare under low illuminative environments. This research firstly studied the literatures to explore the drivers' visual ergonomic as well as the optical properties of LED, and investigated the relatively existing norms for engineering vehicle LED display signs. Three variables were set in this study: three kinds of ambient illumination, four kinds of luminance contrast and two kinds of character form. In the first phase of the experiment, subjects observed LED display signs in both near and distant locations and filled out the SWN scale (Subjective Well-being under Neuroleptics), and in the second phase, subjects were then asked to moved forward and recorded their perceptions of comfort and glare to distance range. The findings demonstrated that, there was no variation in subjective evaluation to display signs with no backgrounds either in the near or distant locations, while to display signs with backgrounds, the subjects perceptions were the farther the distance, the clearer the legibility; higher ambient illumination could effectively reduce observers' glare perception to LED display signs; display signs with backgrounds at the luminance contrast of 3:1 (L max = 3100, L min = 1033 cd / m^2) showed the lowest uncomfortable and glare level to observers. The two forms of character showed no significant variation in affecting observers in terms of the comfort and glare perception.

Keywords: LED display signs, engineering vehicles, legibility, ambient illumination, luminance contrast.

1 Introduction

LED, a light source made from the semiconductor technology, has been widely used in IT products, communication electronics, display panel, traffic signal, and various instrumental displays. Owing to the price drop and improving product features, High-Brightness LED has gradually replaced the traditional LED, in addition, as the emerging markets have used High-Brightness LED directly, the traditional LED is now only partly utilized in signs, lighting, electronic equipments etc., and its market has been extremely shrinked.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 53–62, 2011.

Currently, High-Brightness LEDs are mainly applied to mobile phones, displays, automotives, lightings, signal lights, and other areas. In viewing the application and future forecast, the DisplaySearch pointed out that the outdoor display has been ranked in the top two demands of High-Brightness LED. Therefore, the LED display immediately provides important and dynamic traffic information to people. According to current experience of LED displays application, in order to obtain an ideal display effect outdoor, the brightness has to be over 4000 cd/㎡ [1]; However, in the night, such high brightness might make the observers feel discomfortable and glare. Take the arrow direction traffic light on the rear of national highway engineering vehicle for example, according to the statistics of National Freeway Bureau, there is about six constructional fatal accidents per year, in which three of them usually occurred during the operation of mobile engineering trucks. Although those accidents were not cause by engineering trucks, all seemed to be relevant to the insufficiency of warning signs on the vehicles and related constructional spots, which also showed a high rate of car accident on the highway constructional operation. Therefore, this research aimed to study how to make the brightness of LED display sign on engineering trucks to reach the most comfortable visual distance which allows drivers to have more time to do proper reactions and judgments, as well as avoid causing glare to drivers when they are driving at close range.

Kazunori Munehiro et al. [2] study pointed out that the legible distance of LED guide light in the daytime is father than road marking, while in the mist the glare level will rise with the increased brightness of LED. As this is the situation, an appropriate change of brightness of LED is necessary. Therefore, installing LED on road traffic does improve legibility of drivers. The study of Uchida Kazuhiro et al [3] has indicated that the low legibility during the night is because of the high contrast brought out by too high brightness, while the text and background within a certain contrast can enhance the effect of visual recognition.

On aspect of the application of LED display signs in each country, according to the Highway Construction Traffic Control Manual of Directorate General of Highways, MOTC [4], there are at least four levels of the brightness of LED display signs in which the darkest level must be the half of the brightest, and the intensity of each LED light must not be less than 2cd. .In the manual of the road construction traffic control and safety equipment of Macao Special Administrative Region Transport Bureau [5], it also mentions that the brightness of LED light on the engineering vehicle should be adjusted in accordance with the surrounding illumination. Japan only stipulates two levels of adjustment for the brightness of LED display signs, dividing into day and night luminance in accordance with the difference of color.

From the various regulations for variable message sign and LED signal on engineering vehicles set by each county, it shows that currently there is no uniform provision to the brightness of LED application on engineering truck. Therefore, this research focused on the brightness contrast background between LED variable message sign and LED display panel, conducting the study of visibility under different illumination environments, to investigate how the luminance contrast of LED display sign reached the farthest visible distance under different illumination environments and without creating glare to drivers in close distance.

2 Research Methods and Experimental Design

2.1 Research Methods

The purpose of this study was to explore:

(1) The farthest and comfortable distance of the luminance contrast of LED display sign under different ambient illuminances.
(2) The effect of the character form of LED display signs to subjects' legibility.

Experimental Variables

(1) The luminance contrast (Lc) of LED sign

This experiment used dimmer to control the different brightnesses of LED and LP9221 UNM6 Luminometer to measure the luminance. The luminance contrast between sign and background was divided into four levels, Lc1(L max = 6200 cd/㎡， L min = 0 cd/㎡)、Lc2(L max=6200 cd/㎡， Lmin = 3100 cd/㎡)、Lc3(L max = 6200 cd/㎡， L min = 2066 cd/㎡)、Lc4(L max =3100 cd/㎡， L min = 1033 cd/㎡).

(2) Ambient illumination

This experiment was conducted at a built darkroom in the third floor of the Department of Industrial Design of Tatung University, and simulated the illumination environments around the clock by using three halogen lamps, in which one lamp was connected to dimmer to control the illumination. The variables of environment illumination were set for three classes, including the bright day (30000 Lux), the dark day (5000 Lux) and night (10Lux). LP 9221 S1 Illuminometer was used to measure the illuminance.

(3) Character exhibition

According to the highway construction traffic control manual of Directorate General of Highways, the main directional contents of advance warning arrow sign on engineering warning car are "← -" and "<<<" two styles. The experimental fetched the front part of the arrow signals of "←" and "<" (Figure 1) to proceed the experiment of the interaction between the two characters. And the sign appeared randomly from four different directions of up, down, left and right to avoid the subjects' learning effect.

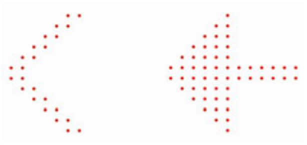


Fig. 1. The characters of arrow sign on LED display panel

Dependent variables

(1) Subjective assessment questionnaire: a questionnaire for the subjects to assess their eyes comfort in observing the LED signs at the nearest distance (9.8M) and the most far distant (57M). Questionnaire items were: 1. I feel eye fatigue, 2. I

think things look difficult, 3. There is a strange feeling around the eyes, 4. I feel numb, 5. I feel headache 6. I feel dizzy 7. This experimental signal is easy to identify, and the five-point scale was applied with point 1 to 5 which represented from "strongly disagree, disagree, fair, agree, strongly agree.", respectively. Items from1 to 6 were negative visual feelings, item 7 was positive feelings to explore the identification degree to sign.

(2) Range of comfort: the distance which subjects feel comfortable while observing the LED display panels – the distance which subject feel the glare while observing the LED display panels = comfort.

(3) Glare range: the distance which subjects feel glare while observing the LED display panel – point of origin = glare range.

2.2 Experimental Equipment

According to the contents of highway construction under traffic control manual, the size of exhibition sign should be in 75cm (L) X 150cm (W); however, due to the limitation of experimental space, the size ratio of the sign in this study was reduced to the size of 30cm (L) × 60cm(W), within a total of 16 × 32 LED lights.

To brightness control, in order to adjust the brightness difference between the middle section (12 x12 lights) of LED display panel and background, the signal part was connected with controller and each LED of 12 × 12 light points was connected to the programmable controller and LUZ-5 relay. The electronic control of signal part was shown in Figure 2.

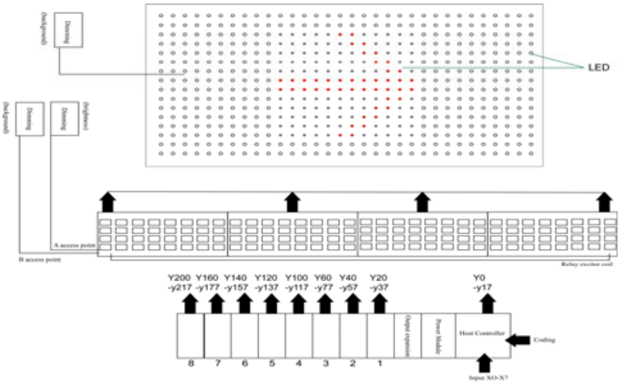


Fig. 2. The diagram of control system of signal section

2.3 Subjects

Before experiment proceeding, this research carried out a visual acuity test to all of the subjects in accordance with the standards of obtaining the automobile driving license specified by the Directorate General of Highways M.O.T.C, that is, the naked vision value of both eyes have to be 0.6 or over and each eye has to be over 0.5 or 0.8 after correction, as well as with field of each eye view reaches 150 degree , and without night blindness as well as also be able to distinguish among red ,yellow and green colors. Based on Tatung University students as sample population, this study randomly selected 29 subjects (15 males, 14 females, aged between 21 to 30 years), while the difference of gender and age were not considered in this experiment.

2.4 Experimental procedure

(1) Subjects firstly filled in the basic personnel information in the classroom and their visual acuity values were confirmed to meet the requirement of this experiment. Next, the researcher presented the experiment details to subjects and explained questions raised by subjects.

(2) Subjects sat in chairs with the same height as the car seat and, under the condition of 2 level of character x 4 level of luminance contrast x 3 level of ambient illuminance of LED display sign luminance and ambient illuminance adjusted by the researcher, observed the LED display panel in both of the nearest distance of 9.8 meters and the farthest distance of 57 meters (Figure 3), and then filled out the subjective assessment questionnaire of their perceptions.

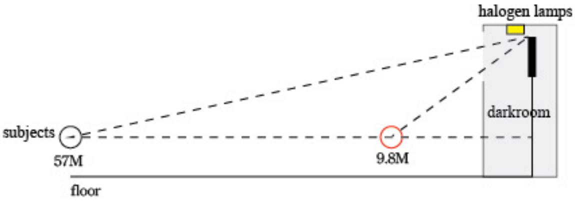


Fig. 3. First phase of experiment

(3) In the second phase, the subjects moved forward slowly from the distance of 57 meters until the they felt comfortable with arrow sign of the LED display panel sand recorded the visual distance at that moment(Figure 4). Then the subjects kept on moving forward until they felt glare to the arrow sign of LED display panel, and again recorded the visual distance at that moment.

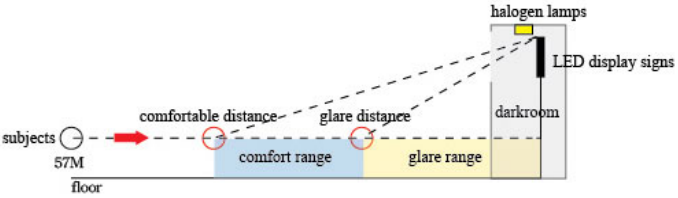


Fig. 4. Second phase of experiment

(4) By the end of each test, the subjects took a short rest to avoid the visual fatigue or discomfort.

(5) Finally, this research applied Microsoft Excel 2007 to calculate the descriptive statistics for all variables of the experimental results in the first stage. The results of second experimental stage were calculated by SPSS 18.0 software to analyze the interaction between the three factors, and if there is not interaction, the One Way ANOVA analysis would be conducted to explore the variation between single-factor variables.

3 Results and Discussion

3.1 The Descriptive Statistics for Subjective Assessment of the Far and Near Distances

The subjective assessments were the assessments of different variables combination of 24 groups at the distance of 9.8 meters and 56 meters. The descriptive statistical analysis of A ~ X groups was shown in Figure 5.

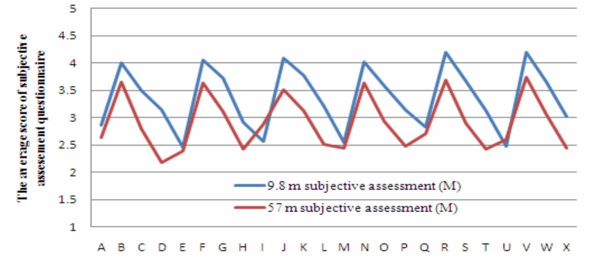


Fig. 5. The subjective assessment of the distances of 9.8m and 57m

Figure 5 showed that, in general, the visual sense in 9.8 m was inferior to 57 m, only the subjective assessments of A, E, I, M, Q, U groups in different distances showed no significant variation, some of them even appeared a phenomenon of the visual sense in 9.8 m was superior to 57m. As shown in Table 1, the statistical visibility compiling from item 7 showed that the variation of average scale of visibility in these six group were less obvious than other groups.

Table 1. The subjective assessment of visibility of A, E, I, M, Q, U groups

Group	9.8m subjective assessment of visibility		57m subjective assessment of visibility	
	Mean	Standard Deviation	Mean	Standard Deviation
A	3.76	.831	3.80	.957
E	4.36	.569	4.04	.935
I	4.36	.757	3.84	1.028
M	4.32	.748	4.36	.860
Q	4.04	1.060	4.04	.790
U	4.20	.866	3.96	.935

This demonstrated that, either in the near or far distances, there was no variation of the discomfort generated by LED display sign without background; while to LED display sign with background, the subjective experience of observing from far distance was superior to LED display signs without background. This finding was in consistent with the conclusion of the study of Uchida Kazuhiro et al. [3] which suggested that, under a certain contrast range, the character and background have the effect of enhancing the legibility.

3.2 Comparison of Three-Factor Multivariate Analysis

The second phase of this research was to conduct experiment for the combination of 3 different variables. In this experiment, the subjects were asked to move forward from the distance of 57 meters to the distance until they felt comfortable to the light source of LED display panel, and then kept on moving forward until to the distance they felt glaze. The subtraction of both distances was the subjects’ comfortable range to their variables combination. In order to understand the significance of variation of ambient illumination, LED luminance contrast and character to the perception of comfort and glare, this research used SPSS18.0 to conduct the three-factor multivariate analysis for collected data, the results showed that, in a comfortable distance range , there were $P = 0.994$, and no significant variation in interaction ($P> 0.05$). In the distance range of glare, the analyzing result showed $P =0.975$, and no significant variation in interaction ($P> 0.05$). To those insignificant variations, the reason we inferred was that, normally the ambient illumination at noon is up to 50000 ~ 100000Lux, while our experimental environment illumination was just 30000Lux.

The interaction between ambient illumination and character form in the range of comfortable distance and the range of glare were significant different with $P = 0.006$ and $P = 0.004$ ($P <0.05$), respectively. The figure 6 and 7 demonstrated that under the 30000 Lux environmental illumination, the range of comfortable distance of arrow sign with tail”←—” was farther than arrow sign without tail ”<”, and the range of glare distance was shorter in arrow sign with tail. It was speculated that under the higher

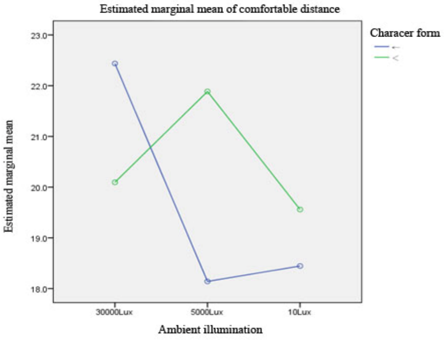


Fig. 6. The comparison of comfortable distances between ambient illumination and character form

ambient illumination, the LED with more display signs offered greater visual effects; while under lower ambient illumination, such as in the evening or night, the glare generated by the LED with more display signs was higher. The interaction between ambient illumination and luminance contrast to the range of comfortable distance and the range of glare distance demonstrated no significant variation with P = 0.772 and P = 0.734 (P> 0.05), respectively. The variation between character form and luminance contrast to the range of comfortable distance and the range of glare distance was also not significant with P=0.940 and P=0.759 (P>0.05), respectively.

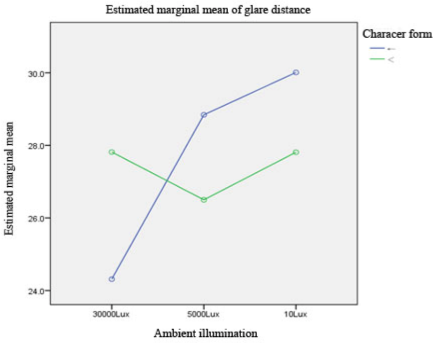


Fig. 7. The comparison of glare distances between ambient illumination and character form

3.3 The Comparison of One-Way ANOVA

Finally, this research preceded the LSD (least significant difference) post hoc analysis to ambient illumination and luminance contrast to explore whether there were variations between the levels of various single factors. The results of individual multiple comparisons were as follows:

(1) The comparison of one-way ANOVA of the ambient illumination

The results showed that, under the environment of 30000 Lux and 10 Lux, the glare distance to observer was significant (P<0.05). The glare distance of ambient illumination at 30000 Lux was short than 10 Lux, which meant that observing under higher ambient illumination would reduce glare extend.

(2) The comparison of one-way ANOVA of Luminance contrast

The LSD multiple comparison of luminance contrast demonstrated that, except Lc1 (Lmax = 6200, Lmin = 0) and Lc = 3 (Lmax = 6200, Lmin = 2066), all variations between groups were significant (P <0.05). The condition of Lc4 (Lmax=3100, Lmin=1033) showed the highest comfort and lowest glare compare to the other groups. Lc3 (Lmax=6200, Lmin=2066) was the second best condition in this analysis. The most uncomfortable and glare condition to subjects was Lc2 (Lmax = 6200, Lmin = 3100). The results showed that the comfortable extend of LED display signs with background were better than those of without background. This result was in line with Uchida Kazuhiro et al [3] finding which concluded that the background brightness can indeed reduce the glare extends generated by the brightness of signs, and lead the observers feel more comfortable.

(3) The comparison of one-way ANOVA of character form

The results showed that there was no significant variation in the comfortable and glare distances (P> 0.05), which indicated that the comfortable and glare distance of subjects' observation to the two character forms were not affected. This represented that there was no difference in observing both of the two character forms applied on the LED of Taiwan engineering vehicles.

4 Conclusion

With the advance of packaging process technology, the controllable factors of luminous efficiency, brightness of LED have also improved significantly. While in pursing the technologic performance as well as applying LED for human observation, it is necessary to take visual ergonomics into consideration. This study investigated whether environmental light illumination would affect the brightness performance of LED display signs. The results of the experimental analysis of this research were as follows:

(1) The distance of LED display sign with background would affect the observer's viewing comfort. In the visible range, the subjective viewing comfort of farther distance was superior to short distance. At the condition of without background (Lmax = 6200, Lmin = 0) conditions, there was no variation of the subjective comfort between observing the LED display panel from the farther or short distance.

(2) The perceived glare extent of observing the LED display sign in the condition of higher ambient light (30000Lux) was lower than the glare extent in the night; on the contrary, subject's viewing comfort in the environment with higher light (30000Lux) was superior to the viewing comfort in the night (10Lux), which indicated that the higher environmental light source around the LED display signs, the lower glare extent the observer perceived.

(3) In the aspect of character form, there was no significant variation between the two character forms in Taiwan engineering vehicle in terms of the extent of comfort and glare. However, in the condition of the interaction with environmental light in 30000Lux, the comfort and glare extent of arrow sign with tail "←" was better than the arrow sign without tail "<"; however, in the night (10Lux) , the above situations were contrary. It was speculated that the ambient light of 30000Lux reduced its value of glare which led the arrow signal with tail to look more comfortable; while in the condition of 10Lux, the perception of less comfort might be because there were more LED lighting points on the arrow sign with tail "←" which resulted the high extent of glare.

(4) From the view point of luminance contrast, similar to LED sign with background on Japan engineering vehicles, the most comfortable luminance contrast of 3:1 was consistent with the results obtained by ANSI / HFS 100 -1988 (1988) recommending the VDT brightness contrast should be 3 at least, while too high brightness will make the observer feel uncomfortable and glare. When Lc = 0, the perception of uncomfortable and glare will be higher than Lc=3, this finding was also in line with the report of enhancing the night visibility of LED sign device provided by Japan Information Processing Juridical Association A which suggested the way of lighting background can prevent the dazzling glare phenomena.

4.1 The Direction of Sequential Studies

(1) The maximum ambient illumination in the experiment of this research was only up to 30000Lux, but the general illumination in the noon is more high at about 50000 ~ 100000Lux, therefore, the ambient illumination should be increased in the sequential studies.

(2) Conduct the experiment in real environment and real size of LED display sign to increase the accuracy.

(3) Include the visibility in the rainfall condition into experiment in accordance with the Taiwan's climatic condition.

(4) Set and explore more luminance contrasts in the experiment in order to obtain the best visibility and comfort.

(5) Target to more character forms to explore different arrow signs (only two character sign in this research).

(6) Aesthetic design should be considered in depth study while in exploring the effective performance of character form.

(7) The future study should consider the legibility of different LED color which was not addressed in this research.

(8) Add the variable of lights flicker frequency to investigate the legible performance.

References

1. LED Industry network (LEDinside) (2007), http://www.ledinside.com.tw/
2. Munehiro, K., et al.: The Monthly Report of Hokkaido Civil Development Institute, No.630 (November 2005)
3. Kazuhiro, U., et al.: Information Processing Society Foundation Research Report, IPSJ (2005)
4. Highway Construction Traffic Control Manual of Directorate General of Highways, MOTC (February 2010)
5. The manual of the road construction traffic control and safety equipment of Macao Special Administrative Region Transport Bureau (January 2009)
6. DisplaySearch, http://www.displaysearch.com.tw/default.aspx

Spatial Tasks on a Large, High-Resolution, Tiled Display: A Male Inferiority in Performance with a Mental Rotation Task

Bernt Ivar Olsen[1], Bruno Laeng [2,3], Kari-Ann Kristiansen[4],
and Gunnar Hartvigsen[1,4]

[1] Department of Computer Science, University of Tromsø, 9037 Tromsø, Norway
[2] Department of Psychology, University of Oslo, Oslo, Norway
[3] Department of Biological & Medical Psychology, University of Bergen, Bergen, Norway
[4] University Hospital of Northern Norway
{bernt-ivar.olsen,gunnar.hartvigsen}@uit.no,
bruno.laeng@psykologi.uio.no, kari-ann.kristiansen@unn.no

Abstract. In previous research we have investigated the effect of screen size on the perceptual mental rotation task (MRT) by comparing performance on a large 230 inches display with that on a standard 14.1 inches laptop display. The former work indicated that females might gain an advantage over males on a larger display. The current study confirms a significant female advantage over male performance in the MRT. However, our current findings helped to reveal that, instead of improving the females' performance, the screen size had a detrimental effect on male performance, while female performance actually remained unaffected by both the large object size than the standard one.

Keywords: Tiled display, Spatial Tasks, Mental Rotation, Sex differences.

1 Introduction

The main motivation behind the present project is to understand the effects on new users of a very large, high-resolution display called a Display Wall [1] in the medical domain of Radiology [2]. The present study extends our previous work where we tested effects of display size as well as effects of *expectations* among the participants [3]. Specifically, we have investigated the effect of screen size on a well-known cognitive task, called "mental rotation", by comparing performance on a large 230 inches display with that on a standard 14.1 inches laptop display. In a previous study, we found that, unexpectedly, women had faster response times (RTs) on the large display than men did. However, the main aim of this previous study was to assess the effect of introducing a novel technology artifact, such as the Display Wall and how the "novelty" of such a situation, compared to the more traditional experience with the much smaller screens of a laptop computer, could affect performance. Hence, in that study, we specifically induced expectations by informing half of the subjects, at the outset of the experiment, that either the Display Wall would yield superior performance or, for the other half of the subjects, that we expected worse performance with the new Display Wall than with the traditional display. We found that females

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 63–71, 2011.

who were told to expect large screen superiority did significantly outperform all other groups in the Display Wall condition. Although suggestive, our previous study was, first of all, based on a rather small sample size and, secondly, it remained unclear whether women would have outperformed men also if they were not given in advance any specific positive expectation. The present study specifically attempted to answer the last question.

1.1 A Couple of Words about the Findings in Our Previous Study

Before we proceed, in order to discuss the present study to full detail, we would like to recap and shed some more light on the results on our first study [3]. The details are available in the previous work, but for ease of reading and understanding we recap some of the results here. The former study had its focus on the effect of introducing a novel technological artifact to subjects who had no prior experience with such equipment. Although we found effects of very large objects deteriorating male performance to the point of a 603ms female advantage in the large screen, there were a couple of concerns that we wanted to address in a new study – in order to "disentangle" the finding of a male disadvantage in mental rotation task with very large objects with that of expectations regarding novel technology.

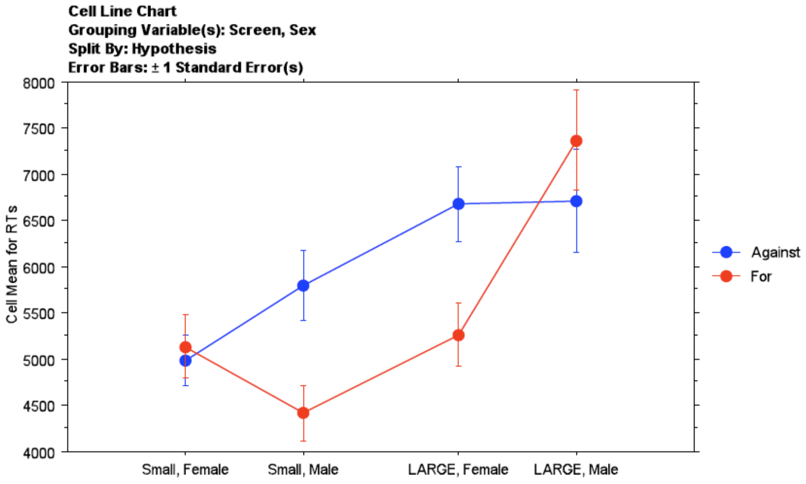


Fig. 1. Mean response times form males/females and large and small display sizes , split by expectation of which condition to perform better in

A sample size of N=40, where participants were effectively distributed among 4 groups (Men with positive/negative "induced expectations" towards the novel display technology and equal setup for females), proved to be too small to yield the expected effects. In Figure 1 we illustrate the most relevant results of that study. The mean response times for each of the groups of participants and the respective experiment conditions revealed that expectation had an influence on male performance only in the small display condition. From these results it seems that males expecting to do better in the large display condition performed better (“for”, “Small, male”) than those believing that the small display was the superior condition. In the Large display condition, females who were expecting the larger display to be the superior condition *did* perform better than females expecting the smaller display to yield superior performance, while males’ performance yielded the opposite result; males expecting better performance in the large display performed slower than those believe the small display to be better. It seems that males performed contrary to “expectations”, while females performed according to “expectations” – at least in the large display condition (no difference in performance between the two groups in the small display condition). In the previous study, hence, “Expectations” (negative/positive) were “forced” upon subjects, while in the current study observations reflect either absence of expectations or what they naturally expect from the novel situation.

Secondly, there was an induced delay in the onset of stimulus in the large display condition, created by network- and processing overhead transferring the screen image from the laptop computer to the computer cluster feeding the projector array. We were not able to measure this delay exactly, which made analysis of within-subjects factors unreliable. This included comparing performance within-subjects on the large-display stimulus to small display-stimulus. To address the issue of delay in the large-display stimulus we have used high-speed camera to measure the mean delay in order to subtract this from the timings of subjects’ performance in the large display condition. Hence, we were able to, this time around, accurately describe how participants generally performed in the mental rotation task on a large display compared to performance on the small display.

2 Method and Experimental Setup

Participants. Thirty-six men and 32 women participated in the study (age range: 18 - 51 years; mean age = 24.0, SD 5.44). All participants participated voluntarily and they were offered two lottery tickets for their participation (this appreciation of their time and effort was introduced after the experiment and, hence, was not used to recruit participants). Nine participants were excluded; four women and five men because they failed to reach a criterion accuracy score of 70% correct in the task; mean accuracy score= 59.6 % and 57.9%, respectively. Two more participants were excluded due to technical failure. The descriptions and analyses shown below consist of responses of the remaining 57 participants. Those who did not understand Norwegian received all the relevant information, instruction and the questionnaire in English. All participants had normal, or corrected-to-normal, eyesight. Participants were recruited from the natural sciences or psychology study programs.

Stimuli and Apparatus. We have used a computerized mental rotation task devised by Peters and Battista [4], which was modeled after the classic mental rotation task introduced by Shepard and Metzler [5]. That is, figures composed of block or cubes are presented as two-dimensional visual images that are constructed of several of such cubes that can be perceived as 3-D figures. As in the Shepard and Metzler paradigm, in the present task participants were shown, in each trial, a pair of cube stimuli where one figure either matched with respect to the other (by applying, mentally, a rotation to the figures) or did not (i.e., the other figure was a mirror image of the other, so that no amount of rotation could make the figures identical). Thus, the participants' task simply consists in deciding, as quickly but as accurately as possible, whether the two figures are the 'same' or 'different'.

The display device used in this experiment was what has become to be known a "display wall", which consists of many projectors that back-project a seemingly coherent picture across the canvas (see Figure 2). Display wall technology is the result of a kind of natural development of the display from a small, single display to today's trend of larger displays and multiple-monitor configurations. With a display wall, testing for the effect of display size is equivalent to testing how the size of an object can affect the task of mental rotation. The present large 'display wall' consists of 28 projectors, back-projecting an image onto a screen surface. There are 7x1024 horizontal pixels and 4x768 vertical pixels, seven by four tiles, on the large screen, a total of approximately 22 million pixels. The physical visible screen size of the large screen is 230 inches. Within the color spectrum there are 22 million pixels of red, green and blue. The small screen setup featured a 14.1 inches screen on a laptop computer, a Dell D600 with a native resolution of 1400 x 1050 pixels and a 24-bit color spectrum.

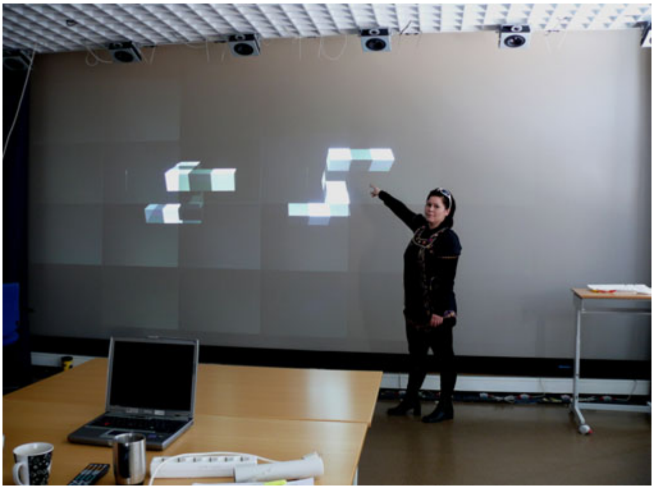


Fig. 2. Mental Rotation stimulus examples on the Display Wall

The mental rotation stimuli where taken from the large stimulus library provided by Peters and Battista [4], which uses wire frame stimuli of the kind introduced by Shepard and Metzler [5]. Stimuli were presented by use of SuperLab® software that was running on the laptop computer, while the image was transferred to the display wall using a 100MB Ethernet interface and a Java implemented display-server running on the Virtual Network Computing (VNC) server on a Dell PowerEdge 2800, with 2 Xeon 3.8GHz/2MB 800FSB, 8GB Dual Rank DDR2 Memory (4x2GB), 146GB SCSI Ultra320 (15,000rpm) 1in 80 pin Hard Drive x 2 with the RedHat Linux operating system. The computer cluster feeding the projectors is comprised of 28+1 Dell 370 PCs with P4 Prescott, 3.2GHz, 2GB RAM, 1Gbit Ethernet and a 48 port HP switch. The SuperLab interface (stimuli) was transferred to the display and enlarged to fit the larger display area of the wall. As a consequence, the number of (perceived) pixels was held constant between the displays, along with the aspect-ration (4:3). As for the screen width and consequential retinal size of the images projected (visual angle of screen), the projected screen (display area covered by Superlab) was measured using a laser-meter to 404cm and 28,5cm for the small screen.

Viewing distance was 370 cm from the large screen, and with the small screen ca. 65 cm. With the laptop computer the participants were instructed to keep a “comfortable viewing distance”. Subsequent to testing, a random sample of the participants was instructed to demonstrate how they solved the task, and measurements were done of the viewing distance. This was recorded and a mean distance of 65 cm was computed. This setup constituted a total visual angle of 57 degrees and 24.7 degrees in the large and small display setup, respectively. This, in turn, constituted between-objects angles of 27.3 and 11.8 degrees, respectively. Total visual angle means the visual angle provided by the display in question, while angle between objects refers to the approximate angle from the person to the midpoints of the objects. During the phase with the large screen, participants used a wireless keyboard in order not to be disturbed by the small screen as they might have been had they used the laptop-keyboard for this task. In the trials with the small screen they used the laptop’s keyboard. The “.” And “Z” keys, which corresponded to a same or different response, were marked with either a green label (for “same”) or a red label (for “different”).

SuperLab©, version 2.02, not only presented all stimuli in a completely randomized order but also kept a record of the key presses and their time occurrence from the onset of the picture. From the Peters and Battista’s stimulus set [4], 19 pairs of images were selected. The second stimulus could differ from the first by 30, 60, 90 120 or 150 degrees of rotation around the vertical axis, for a total 95 pairs. In addition, the second stimulus was either identical to, or a mirror image of the first, for a total set of 190 stimulus pairs. The number of angles was reduced from seven to five from the first experiment to the current. In the de-briefing phase of the previous experiment we had many participants complain that the task was strenuous due to the many trials. This was a consequence of within-subjects design (participants had to complete the experiment in both display size conditions). Excluded angles from the previous experiment were 0 and 180.

In the Large Display condition, the stimuli were transferred from the laptop computer via a 100 Mbit LAN interface (TCP/IP) to a virtual network computer (VNC) server, which would scale the stimuli up to the correct size in order to fill the entire Display Wall canvas. In our previous work [3], we were not able to measure this delay exactly, and given that the variable delay from the previous experiment made analysis within-subjects regarding display size difficult, we attempted to exactly measure the delay in the present study. This was achieved using a high-speed camera, where we videotaped one complete experiment with both displays simultaneously on camera, counted the number of frames that would pass between update of the small screen and subsequent update of the Display Wall screen, recorded the delay between update of the small display and the large display and finally computed the mean delay from a total of 182 samples (trials that were videotaped). What we did in order to measure the delay was to place the laptop display visible in the camera-view so that both displays would be videotaped with 300 frames per second using a Casio Exilim EX-F1 camera for this purpose. Note that during experiments, the laptop computer that was running the experiment software would be set physically aside with the lid closed so that it would not be visible for the participants (i.e., they would only be experiencing stimuli from the large display – even though the laptop computer would still be producing output on the hidden laptop display).

Procedure. The experiment took place in a room containing a "Display Wall" at the Department of Computer Science, University of Tromsø, Norway. Temperature was set at 20° C and light setting to dark. All participants were tested in the same room with the same equipment. Each participant was pseudo-randomly assigned to groups that began testing with the small screen versus the large conditions. Pseudo-random assignment consisted in alternating the conditions to balance the set with sex and first trial-run condition (large or small screen). This was done to counter-balance for the practice effect in within-subjects design [6]. The participants were given 4 training samples before the start of the experiment to ensure that the participant understood the task. The first two training trials included feedback whether the objects were similar/not similar. The task itself was self-paced and each object remained on the screen until the participant made a decision by pressing one of the two keys "." or "Z" to indicate that the shapes were either the same or different. The computer recorded the result for each key press by use of SuperLab© 2.0 software. There were a total of 190 trials for each subject in each of the two screen conditions.

Both small-screen and large-screen conditions took place in the same room, each participant sat at the same table, in the same position, in order to try to keep the environmental variables constant. When both conditions were completed, the participants were given the questionnaire that collected some biographical information, like Sex, age, years of education and type of education. We also included a question to try and record subjective expectations towards which display-condition the participants felt they felt produced better results with regards to speed *and* accuracy. This question was added to the questionnaire in this experiment in order to have something to compare the Expectation variable from the previous experiment with. This time, however, the "Expectation" would be a subjective feeling, rather than something externally produced (by us; the experimenters) *and* recorded *after* the experiments (both large and small screen conditions finished).

Design and statistics. We used a mixed design where Sex (female/male) was the between-subjects factors and Screen (large/small) and Angle (30°, 60°, 90°, 120°, 150°) were within-subject factors. An additional between-subjects factor was Order (large first/small first). Data were analyzed using Statview® (5.0) and SPSS® (v.16).

3 Results

We calculated descriptive statistics for each participant to obtain mean RTs for correct responses and mean % accuracy for each for each combination of variables (Screen Size, Match, Angle). The analyses did not reveal any main effects or interactive effects for Accuracy as factor in the second experiment either. We did have to remove 9 subjects (4 females, 5 males) from the analysis since they scored below a 70% cut-off. Excluded subjects typically had chance performance in one or both screen size conditions. The Large Screen delay was measured on average to be 848ms, which – in the analysis below has been subtracted from the mean Large

Display RTs for all participants. We found a main effect of Screen Size, F(1,55)= 6.69, p= .012, and an interactive effect of Sex with Screen Size, F(1,55) = 6.51, p=.014. The linear relation between angular disparity and RTs were again reproduced (P<0.00, F(1, 4)=77.14). Descriptive statistics confirmed that females were on average 1290ms faster (males: 5347ms; females: 4057ms) than males in the large display condition, while there was no significant difference between the sexes in the small display condition (m: 4208ms; f: 4050ms).

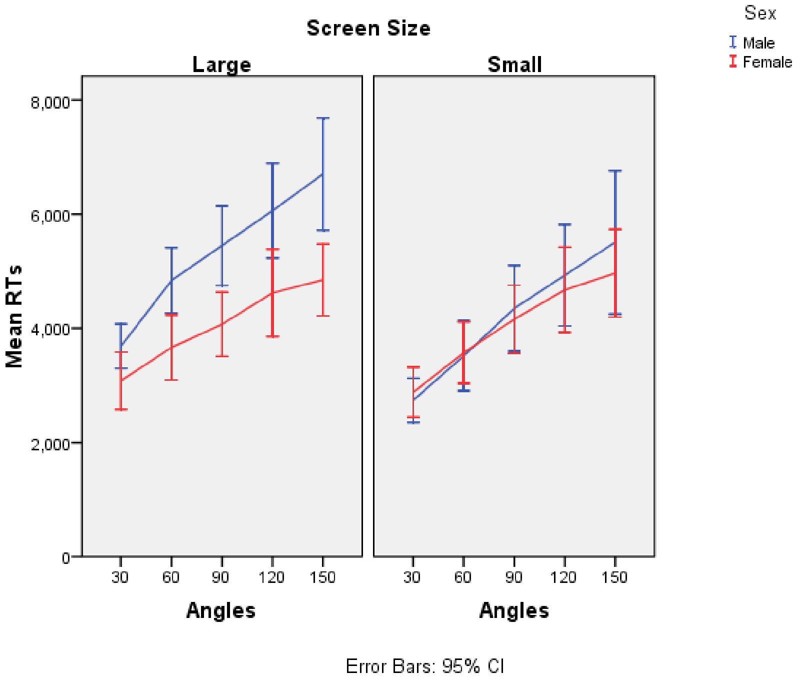


Fig. 3. Mean RTs for the two display conditions for all angular conditions, split by sex. Vertical lines represent 95% confidence intervals for the respective means.

In Figure 3, we observe how the plots of mean RTs for the males and females differ in the large display condition, but do not significantly differ for the normal sized stimulus. The red line in the left half of Figure three (females) follow a curve well below the 95% confidence interval-lines of the blue line (male observed mean RTs with corresponding CIs), while in the regular display (Small) there is no such effect.

As for the Expectation factor, this time around there was no observed interactional-effect of Expectation with Sex in this experiment, as neither Sex (P>0.24) nor Sex * Display Size (P>0.19) showed any significant interaction with Expectation as factor. Even if there was no apparent difference between the genders in their "expectation" towards which screen size they performed better with – in order to investigate participants' expectations towards the displays further, we plotted the mean RTs for each group for all angular conditions of degree of rotation (again: participants responded to whether they thought their performance was better on the large or the small display), presented in Figure 4. What we can see from Figure 4 is that the participants that expected their results to be better on the large display (red line) performed slower than the group that expected their results to be better on the small display.

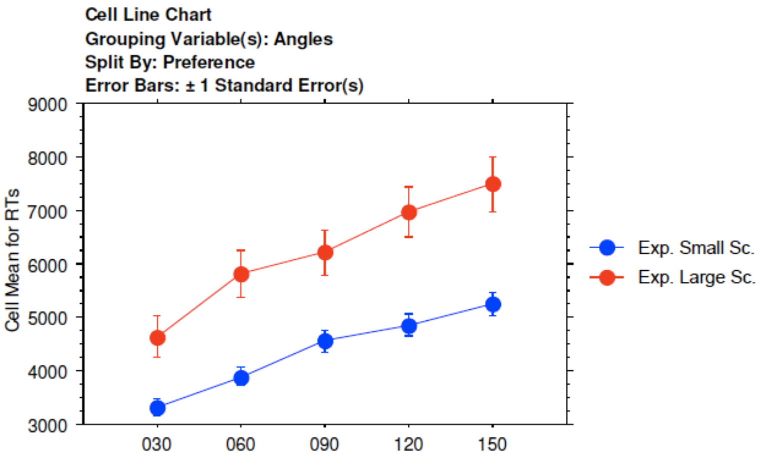


Fig. 4. Experiment 2: mean RTs for different angular conditions, split by which display condition the participants believed their performance was best on

The points on X-axis represent screen size; while points represent observed mean RTs for the different groups with bars representing ±1 SE. We also observe that the positive linear trend for increasing angular rotations is apparent from Figure 4.

4 Discussion

We have confirmed in this replication of our previous study [3] that females outperform males with respect to task efficacy (speed) in the task of mental rotation on large objects (i.e., large displays that cover about 50 degrees of our visual field). Moreover, this effect does not seem to imply females performing faster on larger displays, but rather that the larger (than normal) objects have a detrimental effect on male performance, so as to produce this unusual effect of female superiority on the task of mental rotation.

In our previous work [3] we did observe a female advantage with identical large stimuli as in the current experiment. However, we could not reach a conclusion, based on those data, about whether this female advantage was due to the objects's size or if it was due to expectations alone – or a combination of these factors. From Figure 1 of the present study, it seems that those females who spontaneously expecting to do better with a large display did disproportionally better than those expecting to do worse with a large display than with a normal sized laptop computer display. The current results lead us to believe that Expectation in the previous work had in fact *little or nothing* to do with the "improved" performance in that experiment for females relative to males in the large display condition. In fact, there is no actual "improvement" to be gained from perceiving large objects in the mental rotation task, for either females or males – with respect to either speed or accuracy. What we have observed in this experiment is a significant detrimental effect of large objects on male performance on the mental rotation task, while female performance on average remains virtually unaffected by the larger object size.

Acknowledgements. We would like to thank the Display Wall project and professor Otto Anshus in particular for the use of the equipment and their lab. Our gratitude is also extended to the technical personnel at the Department of Computer Science, University of Tromsø.

References

1. Wallace, G., Anshus, O.J., Bi, P., Chen, H., Chen, Y., Clark, D., Cook, P., Finkelstein, A., Funkhouser, T., Gupta, A., Hibbs, M., Li, K., Liu, Z., Samanta, R., Sukthankar, R., Troyanskaya, O.: Tools and Applications for Large-Scale Display Walls. IEEE Computer Graphics and Applications 25, 24–33 (2005)
2. Olsen, B.I., Dhakal, S.B., Eldevik, O.P., Hasvold, P., Hartvigsen, G.: A large, high resolution tiled display for medical use: experiences from prototyping of a radiology scenario. Studies in health technology and informatics 136, 535–540 (2008)
3. Olsen, B.I., Laeng, B., Kristiansen, K.-A., Hartvigsen, G.: Spatial Tasks on a Large, High-Resolution Tiled Display: Females Mentally Rotate Large Objects Faster Than Men. In: Harris, D. (ed.) EPCE 2009. LNCS, vol. 5639, pp. 233–242. Springer, Heidelberg (2009)
4. Peters, M., Battista, C.: Applications of mental rotation figures of the Shepard and Metzler type and description of a mental rotation stimulus library. Brain and Cognition 66, 260–264 (2008)
5. Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. Science 171, 701–703 (1971)
6. Peters, M.M., Laeng, B.B., Latham, K.K., Jackson, M.M., Zaiyouna, R.R., Richardson, C.C.: A redrawn Vandenberg and Kuse mental rotations test: different versions and factors that affect performance. Brain and Cognition 28, 39–58 (1995)

Modeling Visual Attention for Rule-Based Usability Simulations of Elderly Citizen

Aaron Ruß

DFKI (German Research Center for Artificial Intelligence) GmbH,
Alt-Moabit 91c, 10559 Berlin, Germany
aaron.russ@dfki.de

Abstract. Designing systems for the special interests and needs of older user has become an important subject. However, necessary usability evaluations are time and resource consuming. One way of automation lies in simulating UI use. Since substantial sensory and cognitive age-related effects on the human visual system have been observed, mechanisms of *Visual Attention (VA)* are promising candidates for simulating GUI interactions specific for older users. This article discusses VA mechanisms relevant for simulating age-related effects in GUI interactions. An integration of such mechanisms is discussed on basis of the MeMo workbench, a rule-based approach that uses UI interaction simulations for uncovering usability problems. In the end, simulation of GUI interactions cannot replace human-based usability evaluation, but can provide early feedback for GUI designs, reducing time and resource demands for evaluations. In that, VA provides an instrumental framework for considering age-related effects in simulations of GUI interactions by older users.

Keywords: visual attention, user model, usability simulation, deficit, impairment, rule-based, Monte Carlo simulation.

1 Introduction

In recent years, designing *User Interfaces (UI)* for elderly users has attracted increasing interest, not only within the scientific community but also commercially. The UI design for this user group poses many challenges. Not only are there numerous age-related effects that have to be considered for successful UI design, but the group also shows far greater diversity in their needs and preferences than younger groups. This diversity concerns sensor-motor functions as well as cognitive functions. The group of older users consists of the full spectrum of high functioning users who well into their seventh decade show no or very few signs of cognitive decline, as well as users who very early deteriorate physically and mentally.

Many studies have shown the aging process to negatively affect sensor-, motor- and cognitive functions. For instance, sensor acuity generally decreases as well as strength and accuracy in motor functions. And even when cognitive function can be maintained, its particulars change: for instance, *fluid intelligence* (processing speed, working memory, etc.) generally decreases while *crystalline* increases (knowledge, verbal fluency, etc.).

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 72–81, 2011.

Age-related effects on *visual perception* are of special interest when considering the usability of *Graphical User Interfaces (GUI)*. Generally, evaluating usability is a time and resource consuming process and numerous approaches for automation exist [1]. While model-based evaluations [2] mitigate the problem of the time and resource consuming process of recruiting and conducting user-based evaluations by *simulations*, they, instead, require considerable effort for constructing appropriate *user, system,* and *task models*. Essential differences between the various approaches concern the level of detail of the simulation and, usually depending on that, the effort necessary for creating the models. Similarly, the "balance" is affected, i.e. if more effort has to be invested e.g. in constructing the user versus the system model.

MeMo [3] is a workbench for semi-automatically conducting usability evaluations by simulations, i.e. model-based UI evaluations. This article describes a model for *Visual Attention (VA)* that is partially implemented in the workbench for simulating mechanisms of visual perception. The workbench uses models to simulate *users* solving a *task* by interacting with an *UI*. Additionally, the workbench supports constructing and configuring the models for the *UI*, the *task*, and the *user*. The goal is to provide *Information Technology (IT)* professionals with a tool for usability simulations. According to the knowledge that IT professionals usually possess, the workbench requires high effort for creating the *system model* and moderate to low effort for creating and configuring the *task* and *User Models (UM)*. Basic configuration of the UM can be achieved by specifying attributes (e.g. *visual acuity*) and requires no expert knowledge in cognitive science.

Tasks are assumed to be governed by an information-exchange pattern between user and system. Within this limitation, the UM is designed as task-independent as possible in order to allow its application in usability simulations of different UIs with as little effort as possible. Successful task completion is specified by *conditions*, which allows simulating different solution paths and errors, also referred to as simulation of *beginners-, novice-, exploratory behavior* or *generative* approach [2].

2 Related Work

Most simulation-based automation tools [1] either require a concrete task solution to be specified (*expert simulation*; e.g. CogTool) or are based on a cognitive architecture (e.g. SOAR, EPIC, ACT-R) [2]. Expert simulations allow investigating efficiency (*"how fast is the task solved"*) and effort (e.g. the *learning* effort for tasks). In difference, the focus of MeMo lies on investigating the efficacy (*"[how well] could the task be solved?"*).

While cognitive architectures enable constructing detailed high-fidelity cognitive models that allow rigorous validation against experimental data, their UMs are also highly task-dependent. With MeMo, lower cognitive fidelity is traded for a more task-independent UM within the domain of UI usability evaluation.

In [4], novice user behavior is modeled by two interacting probabilistic state graphs (Markov processes) as models for the view of the (novice) user on the system and the (expert) designer's view. Simulated task errors can occur for mismatching states. However, this requires modelers to specify explicitly both state models and the corresponding probability matrices. *Image Processing Algorithms (IPA)* are used to model visual perception and impairments [5]. VA is simulated using IPA by comparing features of potential focus areas to the target stimulus, selecting the most similar.

There are several semantic approaches for simulating *information seeking behavior* using some form of attention mechanism. Not strictly a VA mechanism, SNIF-ACT uses *satisficing* [2] for determining how to process *information patches* (web pages), i.e. deciding when to continue with the current *patch*, when to follow a link to a new *patch*, or when to return to a previous *patch* [6]. Chanceaux and colleagues model the reading of text-boxes in a web page using *font-size*, *locality*, and an *inhibition of return* mechanism [7]. Several approaches simulate web-browsing using methods of semantic analysis (e.g. [2, 6, 7]). Some use VA methods for calculating the order in which a webpage is evaluated (e.g. [7-10]). Often heuristics for VA are used, reducing UM complexity and the need for resource consuming computations (e.g. [7, 11]).

3 The Simulation Workbench MeMo

MeMo is a workbench supporting semi-automatic usability evaluations by means of simulations [3]. The next paragraphs describe relevant parts for the UM simulation.

System Model. The system model represents the interaction logic of the UI that is under inspection. Generally, for each software application, a new system model has to be constructed. The basic objects of a system model are *UI elements* (e.g. buttons, text fields) that offer interactions to the simulated user (e.g. left click on a button). The process model for the system follows a *state machine* approach: UI elements form *system states* (nodes), which are connected by *transitions* (edges). The transitions represent interactions that the UM can select in the current UI state (e.g. clicking a button).

UI elements can to be annotated with relevant attribute values (e.g. *font size* for labels, *contrast* against the background). The rule-based simulation draws on these attributes for calculating probability distributions which the UM uses to select the next interaction.

Rule-based Simulation. The simulations follow a Monte Carlo approach by repeatedly simulating the UM solving a task using probability distributions for deciding the UM's next actions: for each iteration, a task definition specifies the *starting state* in the UI model as well as *termination conditions* for successful task completion. Beginning with the starting state, the UM calculates *probability distributions* for selecting an *interaction* that is available in this UI state. A simulation step comprises three phases, *perception, information processing, and interaction execution* wherein *perception* and *processing* may be reiterated several times before an interaction is finally selected. The UM selects and executes interactions based on calculated probability distributions, causing the system model to change states. This selection process continues with new UI states until the task's termination conditions are met or the UM "gives up" – e.g. because of a lack of viable interaction options.

When calculating the distributions, the probabilities are manipulated by *rules*. Rules follow a typical IF-THEN schema of *condition* and *consequence*. Notably, multiple rules are applied, if their conditions are satisfied. This allows modeling a UM by iteratively extending a set of rules until the UM is represented by a large set of relatively simple rules. For example, the current rule set comprises about 600 rules, derived from literature analysis, experiments, and consulting usability experts [3].

The simulation result is a set of task solutions. In difference to a single solution, multiple solutions can also reveal unlikely but interesting solutions. *Interesting* in this context means solutions, that are non-optimal or even unsuccessful. The frequency of specific solutions can be interpreted as indicator for their importance. Analyzing which rules have fired and lead to non-optimal task solutions, can readily provide semantic explanations for UM decisions (e.g. rule with condition *"if button label X has small font size ..."*) and in consequence offer critique on how to improve the UI.

User Group Model. The UM represents a *user group* and is exposed in different degrees to the workbench user. The workbench GUI allows direct manipulation of a set of mostly intuitive UM attributes (e.g. *age*, *visual acuity*). A considerable part of the UM is comprised of rules, that are defined using a *XML Schema Definition (XSD).* During simulation, the rules inform probability distributions. For this, the rules draw on UI *features* (UI element attributes that represent their perception by the UM) and UM attributes. Accordingly, different rule sets define different UMs.

Lastly, part of the UM is "hard-wired", implemented as software-modules. For instance, the UM follows the *Model Human Processor (MHP)* approach [2] where each simulation step is comprised of perception, information processing (*cognition*) and interaction execution (*motor*). The ontological commitment to these three phases is implemented in form of software modules.

Task Definition. In task definitions the *conditions* are specified, that determine when a task is successfully completed. Additionally, the *starting state* is specified, i.e. the system state in which the UM starts solving the task. The definition also contains task specifics for the UM, mainly *task knowledge*; the UM employs the specified task knowledge in an information-exchange strategy [3] similar to the *label following* approach [2, 4].

4 Visual Attention for Usability

Visual Attention (VA) is an integral aspect of usability evaluations – be it explicitly or implicitly. In methods considering VA directly, this helps to answer questions about the *If,* the *When,* and *How Easily* users may find task-relevant GUI elements (e.g. [6, 12, 13]). Implicitly, VA plays a role when considering properties that concern visual saliency, as for example *contrast of luminance and color, size, layout, composition, readability,* etc.

In context of automated usability evaluations, considering VA enables the simulation of various related user behaviors for revealing usability problems. For instance, VA allows taking *sequence effects* concerning GUI displays into account: a UM scanning a GUI selects a *sufficiently* fitting GUI element (i.e. matching the task goal), instead of the optimal element that would appear later in the UM's scan path. Such sequence effects can be caused by misperceptions as for example reading errors, or by some misleading (semantic) similarity to the optimal choice. For simulating elderly users, such usage errors become especially interesting, since declining sensory functions may increase perception errors.

4.1 Visual Attention

Currently prevalent, space-based theories of *visual attention (VA)* employ a *spot light* metaphor describing the attention process [2, 14, 15]. Conceptually, at least two important components drive VA: a bottom-up, signal-driven process and a top-down, cognitive process [16]. Their individual impact on VA is highly variable, depending on the visual signal (e.g. if it is a purely random pattern, has some structure, or is even meaningful) as well as the context (e.g. which task the viewer is currently pursuing or if expectations are involved, induced by prior knowledge). Many of the relatively fast bottom-up processes can be approximated as *parallel* working and *pre-attentive* [15].

Saliency maps are a well-known concept in vision models for determining first and successive fixations (i.e. the *scan path*), derived from saliency values computed for the visual scene (e.g. [17]). Biologically inspired, the saliency map architecture is based on *feature integration theory* and has been implemented in several VA models. Commonly, these models analyze *features* of the visual image (e.g. color, direction, movement), resulting in separate feature (or *conspicuity*) maps which then are combined into a conjoint saliency map for the image – some models also explicitly consider top-down influences on saliency (e.g. [14]).

Most models employ a winner-takes-all strategy for determining the point of first fixation, using the most salient region. A scan path is derived by selecting the next most salient areas, where previously fixated areas receive reduced saliency in order to facilitate focusing new regions (*inhibition of return* mechanism, e.g. [7, 18]).

However, the influence of bottom-up processes on visual saliency has been shown to be highly dependent on the task pursued, with predictions most accurate for non-specific viewing of artificially generated displays. For instance, in search tasks, bottom-up saliency can be increasingly overridden by top-down processes or even counteracted [19, 20]. Similarly, meaningful content of an image usually informs top-down influences on saliency. Generally, bottom-up saliency takes more precedence, if the viewer is less familiar with the image content. For instance, [18] describes an experiment, where domain specific images were shown to domain experts and non-experts. Comparing eye movement data with a VA model revealed that non-experts were influenced more by bottom-up saliency, whereas experts focused more on semantically relevant regions. In addition, prior knowledge or task demands can prime saliency of visual features (e.g. search for red objects) [15] as well as determine preference to search by specific strategies or concentrating the search on promising image areas [18].

4.2 Visual Attention and Effects of Ageing

VA is strongly influenced by bottom-up as well as by top-down components. Thus, when considering the effects of ageing on VA, the impact on sensory capacities as well as on cognitive functions are of interest. In this, memory is not only relevant for considering top-down effects on VA, but also for how perceived stimuli are processed. For instance, a known strategy for dealing with the restricted processing capacities is *re-coding* (*chunking*) or *grouping* [21].

Physiologically, visual acuity mainly decreases due to changes to the lens [22]. Additionally, opacity of the lens increases, decreasing the intensity of light passing through and causing reduced contrast perception [23]. Since contrast – i.e. the perception of differences – allows to structure a scene and to distinguish objects, it can be considered the most important bottom-up saliency feature.

Generally, corrected-to-normal eyesight can be maintained well into the sixth decade after which visual acuity declines rapidly [22]. In combination with loss of *contrast sensitivity*, visual acuity is disproportionally exacerbated under conditions of low luminance (and low contrast) [22]. This may have effects on the readability and discrimination of elements in GUI designs.

Visual perception in old age is worse than would be expected from sensory decline alone, which can be explained by exacerbating neuronal and cognitive age-related developments [22-26]. A general explanatory construct for age-related effects is *inhibition control*, i.e. the ability to resist interference. In the visual context, this shows in the form of a decreased resistance for salient distractors [22, 27], i.e. older users may be more easily distracted and misled by visually salient GUI objects that are not task-relevant.

With regard to *memory*, similar developments concerning *speed reduction* and *inhibitory control* (*interference*) have been observed [22]. In general, most *short-term* and *working memory* systems show considerable age-related degradation with the exception of *verbal memory* (i.e. understanding of word meaning) [28]. For GUI design, this means that older users may increasingly face problems when the amount of steps for solving a task increases – especially when combined with the need to memorize information between steps.

Additionally, age-related effects have been identified for different ways of accessing declarative[1] memory (i.e. memory for *facts*): *recollection* exhibits strong age-related effects, i.e. the access of memory that is specific with regard to a certain context. Mostly unaffected by age is memory access by *familiarity*, i.e. the non-specific recall of memory (e.g. general knowledge, such as word meaning, without relating it to some specific occurrence or context). With regard to GUI design, this suggests that older users may have more difficulties to learn the use of GUI elements that function and behave substantially different, depending on context. This may also affect perception and expectations about design and layout regarding *consistency*.

In summary, with increasing age, the sensory capacities for vision are negatively affected – this is exacerbated by cognitive factors. Affected are visual acuity, contrast, and color sensitivity, accompanied by the need for higher light intensities. Cognitive processes are slowed and temporal resolution of perception is decreased [22]. In addition, attention focus is more easily intruded by interfering stimuli, i.e. inhibitory control is compromised. These age-related effects can have a substantial influence on GUI usage and are of special interest when evaluating the usability of GUIs for older users.

[1] Non-declarative memory (e.g. habitualized strategies) exhibits only minor effects [22].

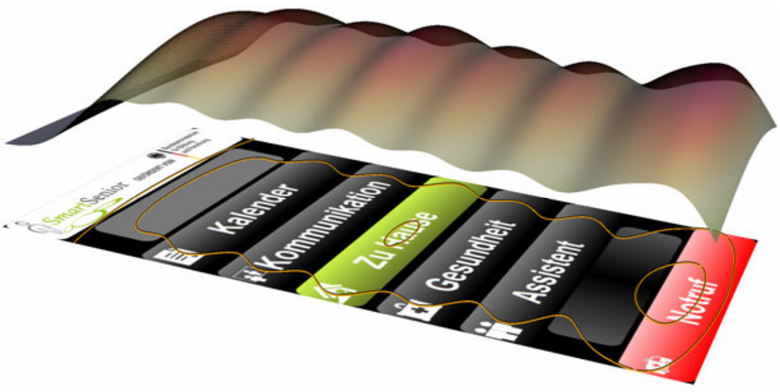


Fig. 1. Bottom-up saliency for GUI elements. The *contour line* renderings implicate two salient areas. The implemented saliency model only calculates "interaction objects" (e.g. buttons), by using Gaussian functions that consider annotated attributes of GUI elements; this includes attributes e.g. for *luminance contrast* (note that *color contrast* is currently not included).

4.3 Visual Attention Model – Current State of Implementation and Next Steps

The VA model is applied in the *perception* phase of a simulation step in the MeMo workbench. In essence, a "spot" in the current GUI state is selected and the corresponding GUI elements made available to the UM's *information processing*. [2]

During this perception-phase a location-based map of the GUI display for bottom-up saliency is calculated (see Fig. 1). Saliency values in this map correspond to (perceivable) UI elements and are calculated by a rule-based approach (see sect. 3), using features of GUI elements as well as user attributes.

For example, during the perception phase, a high contrast of a button increases its saliency; whereas a rule for considering (age-related) visual acuity problems may decrease saliency disproportionally more for UI elements with lower contrast. An advantage of the rule-based approach is that it can provide explanations for the formation of saliency values: after the simulation, an analysis of executed rules can help identify the reasons for the UM decisions and the course of the UM's task solution.

This rule-based approach uses a simplified representation of the GUI (i.e. a *model*) for calculating the visual saliency. This allows simulations even for only roughly sketched GUI drafts in early development stages; but it also requires the construction of a UI model. According to the development stage, the UI model may at first only contain rough layout and type information (e.g. *"button in the upper left corner"*) and in the course of the development process gain more details (e.g. font size of labels, contrast of GUI elements).

[2] The size of the "spot" area depends on parameters, common in usability evaluations: *distance* to the display, *resolution* of the display; as parameter for the visual angle (*fovea*), the EPIC default value 2° is used (e.g. 60 cm *distance* and a 20'' display with 1600x1200 px [100 ppi] would approximately result to a 2x2 cm "spot"). Due to *inhibition of return*, a selected area receives a decreasing saliency reduction in following simulation steps.

The next steps for the implementation are to use spatial information to simulate grouping effects of neighboring elements (*chunking*, see sect. 4.2). This also provides the prerequisite for simulating memory concerning location sensitive information[3]. In the *CODE Theory of Visual Attention (CTVA)*, Laplace distributions are used for modeling proximity effects [9]: nearby elements merge and amplify their saliency in order to explain grouping effects (visual chunking). The effort of attentional focus is represented as a *perception threshold* cutting the "height" of Laplace distributions (e.g. *contour lines* in saliency visualizations, see Fig. 2 and Fig. 1): with low thresholds, elements tend to be perceived as a group, whereas high thresholds allow perceiving individual elements. This provides a basic mechanism for *visual grouping* by proximity. With regard to low-detail-level models ("GUI sketches"), proximity may be the only information available for deriving visual groups. With more design information available, the grouping mechanism can also consider more features.

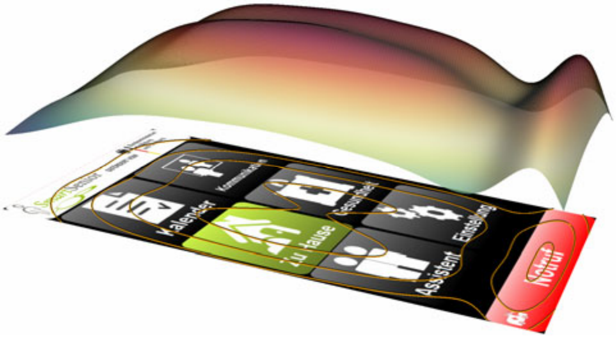


Fig. 2. Alternative layout to Fig. 1. Here, the contour lines suggest differing visual groupings for UI elements in the upper image region, depending on different "perception thresholds".

Furthermore, the grouping information will be exploited for simulating limited cognitive resources: first, the perception threshold is adapted, in order to allow only a limited amount of groups to be inspected at the same time.[4] Then, a selected group is "zoomed in" and the process is repeated as needed: threshold adaption, group selection, zooming in. Depending on "saliency contrast" between GUI elements, different search strategies may be employed, e.g. selecting the most salient (perceivable) group versus systematically scanning groups with similar saliency. This enables simulations of usability problems due to sequence effects (see sect. 4). Additionally, this method allows estimating cognitive workload in terms of attention and memory demands.

[3] E.g. simulating the expectation that a certain GUI element can be found at a specific position or in relation of other "nearby" elements.

[4] The amount of perceivable groups may be influenced by memory limitations (see sect. 4.2).

5 Conclusion

In this article, we considered *Visual Attention (VA)* and its importance for evaluating GUIs. In this regard, *age* has notable effects on VA mechanisms, e.g. loss of sensory acuity, slowing and loss of several cognitive capacities. Accordingly, VA mechanisms provide an expedient framework for incorporating age-related effects in usability simulations.

Using the MeMo workbench, we examined a partially implemented model for bottom-up VA mechanisms, focusing on aspects that are relevant for age-related effects. In this approach, analyzing the formation of rule-based saliency maps can readily provide (semantically relevant) explanations for simulated usability problems.

In conclusion, it is important to note that we are a long way from simulations that can replace human-based usability evaluations. However, they can provide an early and cost effective feedback for UI designs while alleviating the need for extensive usability and cognitive science knowledge on part of the “conductors”.

Acknowledgements. Work presented in this article was funded within the SmartSenior project by the German Federal Ministry of Education and Research (BMBF, FKZ 16KT0902); GUI screenshots with courtesy of Ehsan Gholamsaghaee.

References

[1] Ivory, M., Hearst, M.: The state of the art in automating usability evaluation of user interfaces. ACM Comput. Surv. 33, 470–516 (2001)

[2] Sears, A., Jacko, J.A.: The Human-Computer Interaction Handbook, 2nd edn. Lawrence Erlbaum Associates, Mahwah (2007)

[3] Engelbrecht, K., Kruppa, M., Möller, S., Quade, M.: MeMo workbench for semi-automated usability testing. In: Proc. 9th Interspeech, Australia, pp. 1662–1665 (2008)

[4] Biswas, P., Robinson, P.: Automatic evaluation of assistive interfaces. In: Proc. 13th International IUI, pp. 247–256. ACM, New York (2008)

[5] Biswas, P., Robinson, P.: Modelling perception using image processing algorithms. In: Proc. Human-Computer Interaction, pp. 494–503. British Computer Society (2009)

[6] Fu, W., Pirolli, P.: SNIF-ACT: A cognitive model of user navigation on the World Wide Web. Hum. Comput. Interact. 22, 355–412 (2007)

[7] Chanceaux, M., Guérin-Dugué, A., Lemaire, B., Baccino, T.: A Model to Simulate Web Users’ Eye Movements. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5726, pp. 288–300. Springer, Heidelberg (2009)

[8] Kitajima, M., Polson, P., Blackmon, M.: CoLiDeS and SNIF-ACT: Complementary models for searching and sensemaking on the Web. In: HCIC Winter Workshop (2007)

[9] Pirolli, P., Card, S., Van Der Wege, M.: Visual information foraging in a focus+ context visualization. In: Proc. SIGCHI, pp. 506–513. ACM, New York (2001)

[10] Stone, B., Dennis, S.: Using LSA Semantic Fields to Predict Eye Movement on Web Pages. In: McNamara, D.S., Trafton, J.G. (eds.) Proc. 29th Cognitive Science Society Conference, pp. 665–670. Lawrence Erlbaum Associates, Mahwah (2007)

[11] Halverson, T., Hornof, A.: A minimal model for predicting visual search in human-computer interaction. In: Proc. SIGCHI, pp. 431–434. ACM, New York (2007)

[12] Faraday, P.: Visually critiquing web pages. In: Proc. 6th Conference on Human Factors & the Web, Texas (2000)

[13] Kitajima, M., Blackmon, M., Polson, P.: A comprehension-based model of Web navigation and its application to Web usability analysis. In: Proc. HCI, pp. 357–373 (2000)

[14] Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vis. Res. 45, 205–231 (2005)

[15] Wolfe, J., Horowitz, T.: What attributes guide the deployment of visual attention and how do they do it? Nat. Rev. Neurosci. 5, 495–501 (2004)

[16] Treue, S.: Visual attention: the where, what, how and why of saliency. Curr. Opin. Neurobiol. 13, 428–432 (2003)

[17] Itti, L., Koch, C.: Computational modelling of visual attention. Nat. Rev. Neurosci. 2, 194–203 (2001)

[18] Underwood, G.: Cognitive processes in eye guidance: algorithms for attention in image processing. Cognit. Comput. 1, 64–76 (2009)

[19] Betz, T., Kietzmann, T., Wilming, N., König, P.: Investigating task-dependent top-down effects on overt visual attention. J. Vis. 10ax, 1–14 (2010)

[20] Einhäuser, W., Rutishauser, U., Koch, C.: Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. J. Vis. 8, 1–19 (2008)

[21] Miller, G.: The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information. Psychol. Rev. 63, 81–97 (1956)

[22] Birren, J.E., et al.: Handbook of the Psychology of Aging, 6th edn., vol. 564. Academic Press, Burlington (2006)

[23] Elliott, D., Whitaker, D., MacVeigh, D.: Neural contribution to spatiotemporal contrast sensitivity decline in healthy ageing eyes. Vis. Res. 30, 541–547 (1990)

[24] Higgins, K., Jaffe, M., Caruso, R., Demonasterio, F.: Spatial contrast sensitivity: effects of age, test-retest, and psychophysical method. J. Opt. Soc. Am. 5, 2173–2180 (1988)

[25] Salthouse, T., Hancock, H., Meinz, E., Hambrick, D.: Interrelations of age, visual acuity, and cognitive functioning. J. Gerontol. B Psychol. Sci. Soc. Sci. 51, P317–P330 (1996)

[26] Wright, C., Drasdo, N.: The influence of age on the spatial and temporal contrast sensitivity function. Doc. Ophthalmol. 59, 385–395 (1985)

[27] Munoz, D., Broughton, J., Goldring, J., Armstrong, I.: Age-related performance of human subjects on saccadic eye movement tasks. Exp. Brain. Res. 121, 391–400 (1998)

[28] Park, D.: Ageing and memory: mechanisms underlying age differences in performance. Aust. J. Ageing 17, 69–72 (1998)

Operational Characteristics Related to Memory in Operating Information Devices with Hierarchical Menu

Norikazu Sasaki, Motoki Shino, and Minoru Kamata

The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
nsasaki@sl.t.u-tokyo.ac.jp

Abstract. In an aged society, easy-to-use information devices are necessary. In order to develop information devices which elderly adults can use easily, it is important to bring out characteristics of elderly adults in using information devices based on their cognitive functions. In this study, the authors focus on a relation between memory function and exploration behavior in a hierarchical menu while learning process. An experiment is conducted with eight elderly adults and eight young adults. In this experiment, a hierarchical menu composed based on the library classification, not actual hierarchical menus, is used in order to eliminate differences of knowledge about information devices between elderly and young adults. As a result, it seems that decrease of episodic memory increases a possibility of improper selections in the hierarchical menu when doing the same operations as previous operations.

Keywords: Elderly adults, Hierarchical menu, Memory function.

1 Introduction

Since growth of an information society force the elderly to use complicated information devices such as a mobile phone, it is necessary to develop information devices which elderly adults are able to use easily. Elderly adults feel difficulties in operating devices more than young adults due to decrease of physical and cognitive functions with aging. Therefore, there have been many studies dealing with the decrease and aiming at ease-to-use devices for elderly adults.

In recent years, many studies focus on decrease of cognitive function with aging. Satoru S. et al.[1] investigated the effects of age-related decrease of cognitive functions on use of a ticket vending machine for the Japanese bullet train and indicated that age-related decrease of cognitive functions is differentially related to problems with using information devices. Kartin A. et al.[2] indicated that spatial abilities are essential and an adequate mental model is decisive for PDA navigation performance as a result of examination with elderly and young users.

However, few previous researches, including above researches, focus on continual use of the same device, repeat of the same operation and decrease of memory function. Although taking advantage of experiences of the same operation performed previously is one of important things to use devices easily, elderly adults might be confused about how to reach an operational goal which they have experienced before due to decrease of memory function. Therefore, in order to develop an easy-to-use

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 82–91, 2011.

device, it is important to clarify operational characteristics related to memory function and to design devices based on its characteristics. In this paper as a first step, operational differences between elderly adults and young adults are compared from a perspective of memory function. In order to examine the differences, a hierarchical menu is used because it is a common component in many information devices.

2 Relation between Memory Function and Exploration Behavior

Considering use of common information devices, we repeat the same operation many times. Because users can operate information devices once they learn how to use them through experience of operations, it is important to learn for using information devices. However, during the process of learning how to use, it is required to recall the previous experience for smoother operation when repeating the same operation on the same device. Recalling the previous experience is related to memory function. The objective in this section is to consider effects which decrease of memory function has on repeated exploration behavior in the learning process.

2.1 Relation between Memory Function and Operation of Information Devices

Tulving [3] proposed episodic memory and semantic memory as categories of long term memory. He defined those memories as follows. Episodic memory is “information about temporally dated episodes or events, and temporal-spatial relations among these events”. Semantic memory is “a mental thesaurus, organized knowledge a person possesses about words and other verbal symbols, their meaning and referents, about relations among them, and about rules, formulas, and algorithms for the manipulation of these symbols, concepts, and relations”.

Linton [4] presented a conclusion that the number of trials or experiences has contrastive effects on episodic and semantic memories as a result of six-year study in which memory tests with her own diary was conducted. Although increased experience with any particular event class increases semantic or general knowledge about the event, episodic memory becomes confusable and can not be distinguished. That is, repeating similar experiences transforms episodic memory into semantic memory. Therefore, episodic memory seems to play an important roll in the learning process.

Considering use of information devices, whether users can recall previous operational information from episodic memory affects whether they can use information devices smoothly in the learning process. Recalling which item user selected or how a screen changed leads to smooth operation of information devices.

Fig. 1 shows a relation between memory function and operation which we consider based on the seven stages of user activities by D. A. Norman [5]. Operational process includes estimation, selection and evaluation. Users estimate which item more likely to lead to an objective function, select item based on the estimation and evaluate whether or not the selection is correct. In the evaluating process, users store a result of evaluation in episodic memory. In the estimating process, users estimate based on information in episodic memory if they performed the same operation previously or knowledge in semantic memory if they perform the operation for the first time.

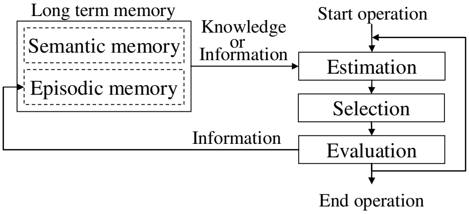


Fig. 1. A Relation between Memory Function and Operation of Information Devices

2.2 A Hypothesis about Effects of Decrease of Memory Function on Learning Process in Operating Information Devices

In the field of cognitive psychology, episodic memory decreases with aging [6]. On the other hand, semantic memory does not decrease with aging [7]. Therefore, we made a following hypothesis. Decrease of episodic memory increases a possibility of improper selections in the hierarchical menu when doing the same selections as previous selections because it makes storing information about operations difficult. Moreover, when selecting an improper item because of failing to recall information from episodic memory, a selected improper item is the same as previous selection because decision of an item to select is related with semantic memory.

3 Experimental Method

The objective of this experiment is to examine the hypothesis about effects of decrease of episodic memory on operating information devices in the learning process. Therefore elderly participants and young participants do exploration tasks in a hierarchical menu. In order to extract the difference in learning process between elderly adults and young adults, participants do the same task four times and tendencies of results are compared among participants.

3.1 Hierarchical Menu Used in This Experiment

In this research, we focus on effects of decrease of episodic memory on operating information devices in the learning process. There is a possibility that use of a hierarchical menu of a mobile phone for this experiment affects exploration behavior of elderly adults and young adults due to differences of knowledge about information devices which they have. Therefore we make a hierarchical menu for this experiment based on the library classification, Nippon Decimal Classification in order to eliminate differences of exploration behavior between elderly adults and young adults based on differences of their knowledge.

A task is to search the hierarchical menu for a designated book. Searching the hierarchical menu in this experiment for a book seems to simulate searching hierarchical menus of information devices for an objective function.

3.2 Participants

Eight elderly adults (four women and four men), ranging in age from 65 to 74 years (M = 69.0, SD = 3.0) and eight young adults (one woman and seven men), ranging in age from 20 to 23 years (M = 21.5, SD = 0.9) participate in the experiment. They use mobile phone usually. We got informed consent from all participants before the experiment.

3.3 Experimental Equipment

The hierarchical menu and the task were made up by Visual C++ 2008 and run on a Dell Latitude D520 notebook PC that was connected to a 22 inches display for PC with a display resolution of 1680 × 1050. Operations required to do the tasks was performed with a computer mouse. In order to measure eye movement of participants, a SMI contact-free eye tracker, iView X^{TM} RED, is set on the display and controlled with Lenovo T500 notebook PC. Fig. 2 shows the appearance of this experiment.

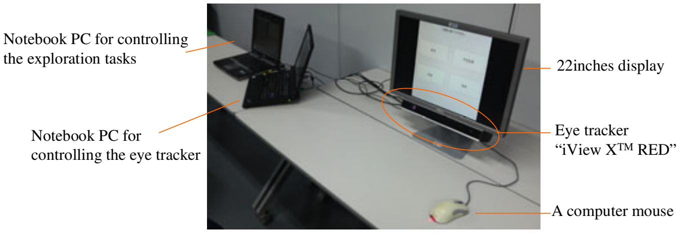

Notebook PC for controlling the exploration tasks
Notebook PC for controlling the eye tracker
22inches display
Eye tracker "iView X^{TM} RED"
A computer mouse

Fig. 2. Appearance of the experiment

3.4 Preliminary Confirmation

In this experiment, participants have to read letters on the display. If it is difficult for them to read letters, their exploration behavior is likely to be affected by the difficulties. Therefore, as a preliminary confirmation, participants read aloud some sentences composed of letters which are the same size as letters used for the hierarchical menu on the display and we confirmed that participants have no problem reading letters in this experimental condition.

Since operations in this experiment are performed with a computer mouse, participants practice to click buttons on the display in order to get used to perform with the computer mouse. Moreover, we confirmed that participants have no problem using the mouse computer.

3.5 Tasks

Participants search the hierarchical menu previously described for a designated book. One task is to search for one book. Books which participants search for in this experiment are decided under following four conditions: (1) the number of letters composing a book title ranges from seven to nine, (2) the number of letters composing a book title which is converted into only hiragana characters ranges from 10 to 14, (3) book titles do not include words used in the highest layer or second layer of the hierarchical menu, (4) arrangement of books to search for is balanced in the hierarchical menu in order that participants encounter many items evenly.

Fig. 3 shows an explanation of order of tasks. Participants search for four books four times and 16 books once respectively in a total of 32 tasks. Participants search for four books which are searched for four times repeatedly and four books which are searched for once through the tasks by turns. Moreover order of tasks is randomized for each participant. In addition, since point of focus in this study is learning process, we analyze results of tasks related to books searched for repeatedly.

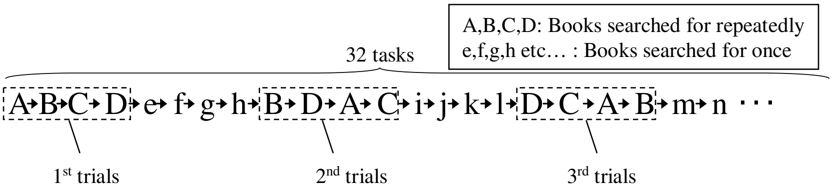

A,B,C,D: Books searched for repeatedly
e,f,g,h etc… : Books searched for once
32 tasks
A→B→C→D→e→f→g→h→B→D→A→C→i→j→k→l→D→C→A→B→m→n · · ·
1st trials
2nd trials
3rd trials

Fig. 3. Explanation of Order of Tasks

Fig. 4 shows screenshots which are used in this experiment. After the confirmation of the book title, participants click on a button placed in the center of a screen and start a task. Following the click on the button, the screen changes into a screen of the highest layer in the hierarchical menu whose screenshot is provided on the left side of Fig. 4. Since the title of the book is displayed at the top of the screen consistently, participants can confirm the title anytime. In layers other than the highest layer, a button, "Return to next superior layer", is displayed at the bottom of the screen. Participants can return to next superior layer by clicking on this button. Forth layer includes various titles of books which are put in double parentheses. When participants select the designated book in the forth layer, the task ends. Before starting exploration tasks, participants search the same hierarchical menu for four books in order to confirm that participants understand the task.

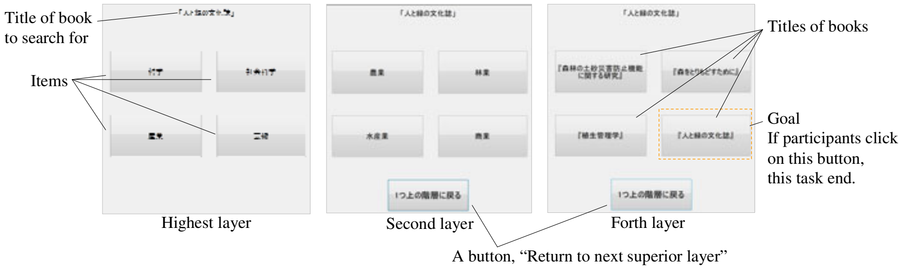
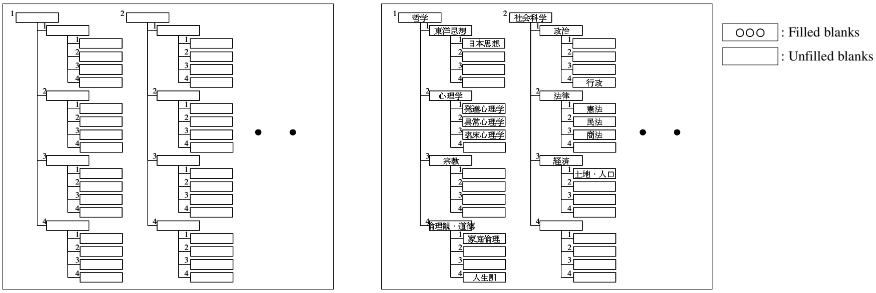


Fig. 4. Screenshots of the Hierarchical Menu in the Experiment

If participants are not in a proper node in second layer to achieve the task within 90 seconds from the start of the task, an experimenter tells participants about the item to select in the highest layer as a first clue. In addition, if participants are not in a proper node in third layer to achieve the task within 150 seconds form the start of the task, an experimenter tells participants about the item to select in second layer as a second clue. These clues are given so that participants reach to the goal after exploring the menu and use the experience of the previous exploration because the point of this experiment is recalling previous selection in the menu.

3.6 Measurement of Episodic Memory

We conduct the fill-in-the-blank question concerning the hierarchical menu in order to measure episodic memory of participants. Fig. 5 shows an appearance of a form used for fill-in-the-blank question. Since measurement of episodic memory is conducted by recalling words showed by experimenter orally or descriptively, in this experiment, measurement of episodic memory is conducted by filling blanks in a form with names of items in the hierarchical menu as shown in Fig. 5. Episodic memories of participants are compared based on the number of filled blanks. Participants do the fill-in-the-blank question five minutes after finishing the exploration tasks in order to prevent recency effect. That is, if the question is conducted immediately after the exploration task, the number of filled blanks increases because of using information existing in short term memory.



Before filling After filling

Fig. 5. Fill-in-the-blank Question concerning the Hierarchical Menu

4 Results and Discussion

4.1 Difference of Episodic Memory between Elderly and Young Participants

Fig. 6 shows a result of fill-in-the-blank question concerning the hierarchical menu in terms of the number of blanks filled correctly. Mann-Whitney U test result showed that there is a significant difference between result of the elderly group and the young group ($p<.01$). Therefore, episodic memory of the elderly group is lower than that of the young group.

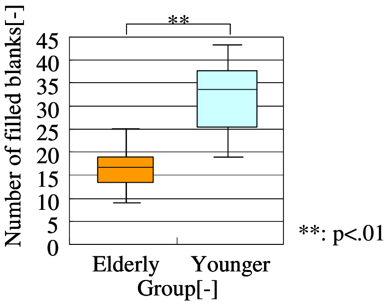


Fig. 6. A Result of the Fill-in-the-blank Question concerning the Hierarchical Menu

4.2 Comparison of the Number of Futile Selections

We compared the elderly group with the young group by the number of futile selections. A futile selection is an unnecessary selection to achieve a task.

Fig. 7 shows results of the number of futile selections with respect to the trial number. Since the number of participants is not large, these results show not average value but median value. Mann-Whitney U test results showed that there is a significant difference between results of the elderly group and the young group in third trials (p<.05).

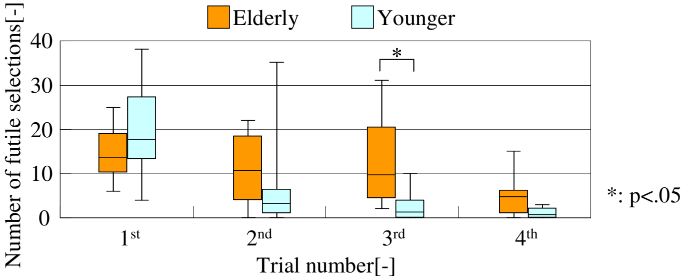


Fig. 7. Results of the Number of Futile Selections

In this experiment, participants search the hierarchical menu composed based on the library classification for books. In first trials, because participants depend on not knowledge about information devices such as mobile phones but knowledge about words, the number of futile selections reflects the difference of knowledge about words among participants. Moreover, because knowledge about words is stored in semantic memory, in first trials, semantic memory, not episodic memory, seems to be related to exploration behavior in the hierarchical menu. In contrast, in second, third and forth trials, it is important to recall what they did in previous trials. That is, it is important to retrieve information from episodic memory. If they can recall which item led to the goal or which item did not lead to the goal, the number of futile selections decreases in second, third or forth trials.

In first trials, there is not much difference between results of elderly and young group. However, after first trials, there are different tendencies between results of the elderly and the young group.

A result of first trial implies that there is not much difference between knowledge of words of the elderly group and the young group. On the other hand, different tendencies of results in trials after second trials, which seems to be attributed to the difference of episodic memory, implies that decrease of episodic memory increases the possibility of improper selections in the hierarchical menu when doing the same task repeatedly.

4.3 Consideration in Elderly Participants

Fig. 8 and fig. 9 show transitions of the number of futile selections in elderly participants and young participants. Participants are put in descending order of results of fill-in-the-blank question concerning the hierarchical menu from left to right. The result of Ss8 is better than that of YSs5 and the results of Ss2 and Ss5 are as good as that of YSs2.

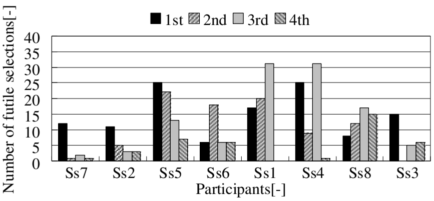


Fig. 8. Results of the Number of Futile Selects among Elderly Participants

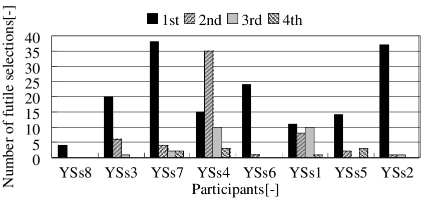


Fig. 9. Results of the Number of Futile Selections among Young Participants

If participants store information about the hierarchical menu gained by performing tasks in episodic memory and use its information, the number of futile selections should decrease with increase in the number of trials. In young participants whose results of fill-in-the-blank question are not lower than those of elderly participants, the number of futile selections tend to decrease from first trials to forth trials except of YSs4. On the other hand, in Ss7, Ss2 and Ss5 whose results of fill-in-the-blank question are as good as those of young participants whose results are relatively low, the number of futile selections tend to decrease from first trials to forth trials. However, in other elderly participants, the number of futile selections does not decrease or decreases once and then increases again even with increase in the number of trials.

These results seem to strengthen the opinion that decrease of episodic memory increases the possibility of improper selections in the hierarchical menu when doing the same task repeatedly.

4.4 Comparison of the Number of Gazed Items

Fig. 10 shows results of the number of items gazed in the highest layer. Because the number of items in the highest layer is four, a maximal value of this index is four. In addition, because the measurement of eye movement of two young participants failed, the results of the young group include results of six young participants. Mann-Whitney U test results showed that there is a significant difference between results of the elderly group and the young group in all trials (p<.05).

If participants know the item to select or arrangement of items, the number of gazed items decreases. Although results of the elderly group and the young group are both decreasing with an increase in trial number, the decreasing trend of the young group is more sharply. This result implies that decrease of episodic memory has an effect on remembering the item to select or arrangement of items and makes exploration in the hierarchical menu difficult.

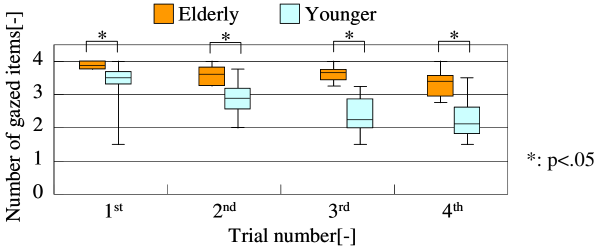


Fig. 10. Results of the number of items gazed in the highest layer

4.5 Analysis of Improper Selections

In order to consider the effect of decrease of episodic memory, repetitional mistakes are analyzed. A repetitional mistake means to select an improper item which participants selected in previous tasks at the node leading to the goal. For example, when a participant selects an improper item, “item X”, in first trials, if he/she selects “item X” in second and third trials, the number of repetitional mistakes counts as two. But even if he or she selects “item X” many times in the same trials, the number of repetitional mistakes counts as one.

If the hypothesis that people decide based on information in semantic memory when failing to recall information from episodic memory is valid, people make similar mistakes. Therefore, repetitional mistakes were analyzed. In the elderly group, 53 of total 79 improper selections made by all elderly participants were repetitional mistakes. The proportion of repetitional mistakes to improper selections is high. This result implies that participants decided an item to select based on knowledge stored in semantic memory because he/she could not recall information from episodic memory.

5 Conclusion

This paper analyzed relation between episodic memory and learning process with the hierarchical menu composed based on the library classification. The major conclusions to be drawn from results are as follows.

- Decrease of episodic memory increases the possibility of improper selections in the hierarchical menu when doing the same selections as previous selections.
- The same improper selections as previous improper selections account for the large portion of improper selections.

In the future work, we are going to examine easy-to-recall factors of information devices, cognitive functions which affect learning, and propose information devices based on gained results. For example, storing information and presenting information about mistakes when users are confused is one of conceivable assistance for users with decreased episodic memory. Because they are more likely to make similar mistakes, it is helpful for them to present information related to previous mistakes.

Acknowledgments. The author Norikazu Sasaki was supported through the Global COE Program, "Global Center of Excellence for Mechanical Systems Innovation," by the Ministry of Education, Culture, Sports, Science and Technology.

References

1. Satoru, S., Takatsune, K.: Effects of age-related decline of visual attention, working memory and planning functions on use of IT-equipment. Japanese Psychological Research 52, 201–215 (2010)
2. Kartin, A., Martina, Z.: Effects of age, cognitive, and personal factors on PDA menu navigation performance. Behavior and Information Technology 28, 251–268 (2009)
3. Endel, T.: Episodic and Semantic Memory. In: Endel, T., Wayne, D. (eds.) Organization of Memory, pp. 381–403. Academic Press, New York and London (1972)
4. Marigold, L.: Transformation of Memory in Everyday Life. In: Ulric, N. (ed.) Memory observed: Remembering in natural contexts, pp. 77–91. W.H.Freeman, San Francisco (1982)
5. Donald, A.N.: Cognitive Engineering. In: Donald, A.N., Stephen, W.D. (eds.) User Centered Systems Design: New Perspectives in Human-Computer Interaction, pp. 31–61. Lawrence Erlbaum Associates, Hillsdale (1986)
6. Lars, G.N., Lars, B., Karin, E., Lars, N., Rolf, A., Gösta, B., Stig, K., Maud, W., Bengt, W.: The Betula Prospective Cohort Study: Memory, Health and Aging. Aging, Neuropsychology and Cognition 4, 1–32 (1997)
7. Marilyn, S.A., Hope, S.H., William, M.: Changes in Naming Ability With Age. Psychology and Aging 3, 173–178 (1988)

Effects of Paper on Page Turning: Comparison of Paper and Electronic Media in Reading Documents with Endnotes

Hirohito Shibata and Kengo Omura

Research and Technology Group, Fuji Xerox Co. Ltd.
6-1 Minatomirai, Nishi-ku, Yokohama, Kanagawa, 220-8668, Japan
{hirohito.shibata,kengo.omura}@fujixerox.co.jp

Abstract. This study compares the performances of paper and electronic media during a reading task that includes frequent page turning. In the experiment, 18 subjects read multi-page documents aloud while referring to endnotes using paper, a large display, and a small display. Results revealed that reading from paper was 6.8% faster than reading from a large electronic display and 11.4% faster than reading from a small electronic display. No difference was found between scores of recognition tests of important words of documents among the three conditions, which indicates that paper is the most effective medium for people to read text speedily without reducing comprehension. Detailed analyses of the reading process show that, in the Paper condition, people perform both text reading and page-turning simultaneously. However, when using computer displays, reading and turning pages were divided completely and performed separately.

Keywords: paper, reading, page turning.

1 Introduction

This study performed a quantitative comparison of the performance of reading multi-page documents with endnotes between paper and electronic media.

Since the 1980s, many experiments have been conducted to evaluate the readability of text documents on paper and displays [1–10]. They mainly focused on presentation properties such as the resolution of a medium, the size of a medium, text font, and the document format, and examined how the difference of these properties affect reading performance. In addition, their experiments addressed sequential reading, whose processes include less movement back and forth between pages of a document or among documents.

However, as Sellen and Harper [11] described based on an observation of real-world reading, only rarely did workers read documents sequentially from beginning to end with turning pages one-by-one, at least in work situations. Readers often refer to the table of contents or references frequently, skim documents, and move back and forth between pages repeatedly while reading. O'Hara and Sellen [12] observed the reading processes used for scientific articles and reported that readers frequently navigated among pages of a document or among multiple documents. They skim

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 92–101, 2011.

articles to grasp the flow of text, move to previous pages to confirm definitions of terms, and turn pages to refer to figures or tables. Additionally, O'Hara et al. [13] observed the professional writing process and reported that writers move attention across documents, lay out documents spatially, and freely annotate documents during writing. They also described that physical paper, as a material artifact, supported these operations effectively.

However, the effect of operability in such operations for reading has been only scarcely investigated quantitatively. This study specifically examines the operation of page turning and quantitatively compares reading performance with frequent page turning using paper and electronic media.

In electronic media, various user interface tools can be used to switch pages, such as scrolls, page-turning buttons (e.g., "previous buttons" and "next buttons"), overview using thumbnails of pages (e.g. the "page navigation panel" of Adobe Reader), physical buttons [14], and gestures [15, 16]. Among them, we specifically examine the operation of scrolling and page-turning buttons because they are common and widely used features that are used to switch pages. Some studies have compared reading performances of scrolls and page-turning buttons [6, 17, 18]. However, the studies were aimed at improving the user interfaces of electronic environments and therefore compared the different user interfaces of electronic media. In contrast, the present study compares reading performance achieved when using paper and electronic media.

2 Hypotheses

As described earlier, in the reading of work situations, readers frequently move back and forth between pages. The aim of the experiment is to analyze how the operability of a medium affects reading performance. For the experiment, we use multi-page documents with endnotes as reading materials. We require subjects to read aloud to follow where the subjects read in documents at each time of reading. By analyzing the discontinuity of reading aloud associated with page turning, we can separate the reading of the main text only, which does not include referral to notes listed on another page, from the reading of whole documents that accompany repeated page turning. Furthermore, we can measure the time used to refer to notes and the time used to turn back to the main text. Hypotheses examined using this experiment are the following four.

First, if reading processes do not include page turning, then we expect that reading from paper is as fast as reading from computer displays. According to an experiment conducted by Gould et al. [5] in the late 1980s, no difference exists between reading speeds attained when reading from paper and when reading from CRT displays. Although the display used in their experiment was a high-performance one in those days (the display resolution was 1024 × 1024), the performance of current up-to-date TFT displays is better than the display that Gould et al. used in their experiment. Consequently, it is apparently natural that the reading speed from displays does not differ from the reading speed achieved when using paper.

Second, if reading accompanies frequent page turning, then we expect that reading from paper is faster than reading from a computer display. Previous observational studies [11, 13, 19] show that paper is preferred for use in the reading of work situations. Adler et al. [20] categorized work-related reading as 10 kinds. Among them, frequently observed reading was reading for cross-referencing, reading to search for answers to questions, reading to support discussion, and skimming. In such reading, people often must turn back and forth between pages. We think that one reason why people prefer paper in work-related reading is that paper supports such reading with frequent page turning.

Third, we expect that people can understand the contents of documents more deeply when reading from paper than when reading from a display. The difference of operability between media is attributable to the difference of cognitive load. During reading with frequent page turning, the cognitive load is less when reading from paper than it is when reading from displays. We expect that this load affects the degree of document comprehension.

Fourth, we hypothesize that people can read documents with endnotes rapidly when reading from paper because people perform both reading and turning pages simultaneously. When reading from paper, people can switch pages with the feeling of hands without using vision. Consequently, people need not look away from the text of documents while turning pages. However, when reading from displays, people must look away from document text to use page turning buttons or scroll bars. We think that this makes it difficult to read and turn pages simultaneously using PC displays.

3 Method

Design and subjects. The experimental design was a 3×2 within-subjects design. The first factor was media (paper, a large display, and a small display); the second factor was parts of reading (reading without page turning and reading with page turning). Each subject performed all conditions of tasks. They performed two trials in each condition. The order of the media and types of reading in the series of subjects' trials were counterbalanced to cancel the effects of the trial order overall.

Subjects were 18 people (9 male, 9 female). Their ages were 20–39 years (avg. 29.1). Each had three or more years' experience of using a PC. The power of vision of each after correction was better than 0.7.

Materials. We created documents for the experiment based on columns of a Japanese newspaper. The documents consisted of two pages: The first page was a main text with eight annotation numbers; the second page was a list of annotation notes. The annotation notes were originally created based on the contents of dictionaries and Wikipedia. We created six documents for the experiment and two documents for a training session. For the six documents used in the experiment, the average length of the main text of the first page was 622.3 characters, and the average length of annotation notes of the second page was 246.8 characters.

Procedure. The task of the experiment was to read the documents aloud. When reading, subjects were required to move to annotation notes immediately after reading words with annotation numbers (*referring*). Subjects were required to return to the former position in the main text and commence reading after reading the annotation notes (*returning*), as presented in Fig. 1.

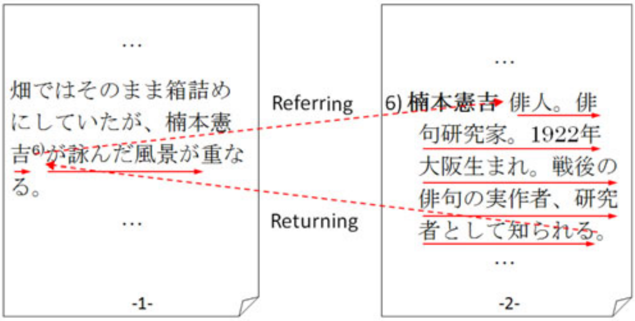


Fig. 1. How to read text documents

We videotaped the images and sounds produced while subjects were reading. To measure the time for referring and returning by specifying the start and the end of continuous sound of voice, we required subjects to read aloud. It might seem that the task setting was artificial because people usually read text without uttering a word. However, a previous study shows that a strong positive correlation exists between silent and oral reading speed: those who speedily read aloud also read rapidly in silence [21].

When reading from paper (Paper condition), documents were printed on one side of B5 paper in black and white and stapled in the upper left corner. When reading from the large display (LD condition), documents were displayed in a 20.1-inch TFT monitor (Diamondcrysta RDT201L, 1600×1200; Mitsubishi Electric Corp.). When reading from the small display (SD condition), documents were displayed in a 10.4-inch TFT monitor (Let's note CF-R3, 1024×768; Panasonic). In electronic conditions (LD condition and SD condition), the OS was Windows XP (Microsoft Corp.). Electronic documents were all PDF format and displayed with Adobe Reader 9 (Adobe Systems Inc.).

In the Paper condition, subjects turned pages using both hands. In the electronic conditions, subjects turned pages using scroll bars or page-turning buttons.

We adjusted the character size of electronic documents to be the same size as those of paper documents. We prohibited changing of the size of displayed characters. In this situation, a single whole page was displayed in the LD condition and about half of a page was displayed in the SD condition. Prior to the experiment, subjects adjusted the position of displays and display preferences to their preference.

We instructed subjects to read documents with comprehension of the contents of the documents. We encouraged subjects to read at a natural speed. However, we required fast and smooth reading when referring to notes and returning to the main text to the greatest extent possible.

After reading all documents, we conducted a recognition test of important words of documents to check how well subjects remembered brief summaries of documents. We did not announce this test for subjects in advance. In the test, for each document, we presented 16 words selected from the document and 16 words as distracters, which were not used in the documents. We required subjects to answer whether or not given words occurred in the documents. In the process of selecting the 16 words, at first two experimenters separately selected 20 important words for each document. The agreement rate was 55.2%. Finally, we selected 16 important words for each document based on discussion between the two experimenters.

4 Results and Discussion

Fig. 2 presents comparison of the reading speed (the number of characters that subjects read per minute) in each condition. The time required for reading the main text was calculated by eliminating the time from the end of reading aloud before referring to notes to the start of reading aloud after returning to the main text. The timing of discontinuity of reading aloud was identified by visual judgment using audio waveform presented by Windows Movie Maker (Microsoft Corp.).

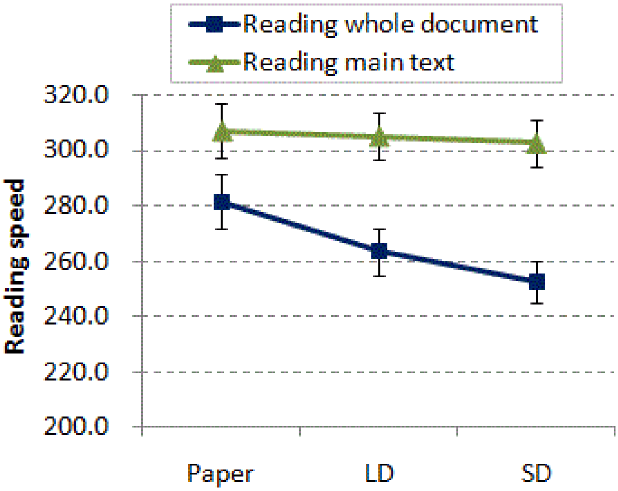


Fig. 2. Reading speed in each condition. Reading of whole documents includes page turning and reading of main text only does not include page turning.

A two-way repeated measures analysis of variance was conducted to assess the reading speed according to the medium used (Paper, LD, and SD) and parts of reading (whole documents and only main text). Results show that the main effects of the medium [$F(2, 34)=5.08$, $p<0.05$] and the parts of reading [$F(1, 17)=151.17$, $p<0.001$] were significant. Interaction of the two factors was significant [$F(2, 34)=11.37$, $p<0.001$]. Then we tested the simple main effects of the medium for each part of reading. Although the simple main effect was significant for the reading of the whole document [$F(2, 34)=15.57$, $p<0.001$], the simple main effect was not significant for the reading of only the main text [$p>0.1$]. According to multiple comparison using the LSD method, reading in the Paper condition was significantly faster than the reading in the LD condition, and the reading in the LD condition was also significantly faster than the reading in the SD condition [$p<0.05$]. Regarding reading documents with frequent page turning, reading from paper was 6.8% faster than reading from large displays and 11.4% faster than reading from small displays.

Regarding the recognition test of important words, no significant difference was found in the scores for different media, which indicates that paper is a medium that enables rapid reading without deterioration in the level of understanding. Our second hypothesis was not supported in the experiment. However, the relation between reading speed and the level of understanding is complementary; subjects might have adjusted the speed of reading to achieve a certain degree of understanding.

No significant difference was found for the main text reading speed. Therefore, reading speed is independent of the medium if reading does not include page turning. However, reading from paper was faster than reading from the displays (LD and SD) in this experiment because the act of referring and returning is performed more rapidly in the Paper condition than in the LD and SD conditions.

Fig. 3 presents a comparison of the processing time of referring to notes and returning to the main text in each condition.

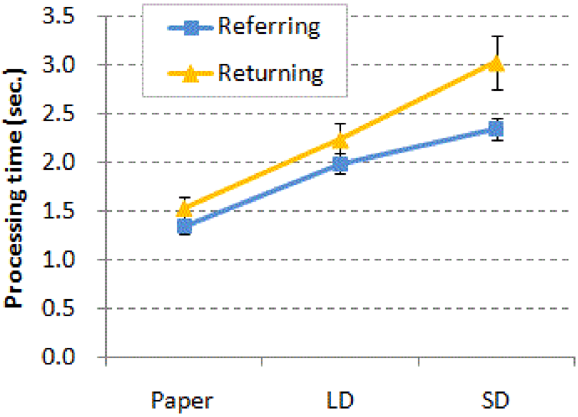


Fig. 3. Time for referring notes and returning to the main text

A two-way repeated-measures analysis of variance was conducted to assess the reading speed according to the medium used (Paper, LD, or SD) and the direction of referring (referring and returning). Results show that the main effects of the medium [$F(2, 34)=69.75$, $p<0.001$] and the direction of referring [$F(1, 17)=9.95$, $p<0.01$]. Interaction of the two factors was also significant [$F(2, 34)=7.426$, $p<0.01$]. Then we tested the simple main effects of the medium for each direction of referring and found that simple main effects were significant for both referring and returning [$F(2, 34)=83.24$, $p<0.001$; $F(2, 34)=42.98$, $p<0.001$]. According to multiple comparison using the LSD method, in both cases, the processing time of the Paper condition was significantly shorter than in the case of the LD condition and the processing time of the LD condition was significantly shorter than in the case of the SD condition [$p<0.05$].

Referring to notes using paper was 32.7% faster than in the case of large displays that can display a whole single page, and 43.0% faster than in the case of small displays that can display half of a single page. Furthermore, returning to the main text from notes using paper is 31.3% faster than in the case of large displays, and 49.2% faster than in the case of small displays.

Turning back to the main text from notes takes more time than referring to notes in all conditions. A reason for this is apparently that it takes time to find the start position of reading after returning to the main text. In the second page of documents, all annotation numbers and annotation words are positioned to the left of lists. Therefore, when referring to notes, people can find the start position of reading easily. However, annotation numbers in the main text are small and scattered in text. Therefore, when returning to the main text, people seem to have difficulty finding the start position of reading.

Fig. 4 presents a comparison of the percentage of page turning in which subjects had started to turn pages before finishing reading the text, where we call these phenomena *simultaneous operations*. For each referring and returning, we judged whether or not two actions of turning pages and reading were performed simultaneously by analyzing the videotapes.

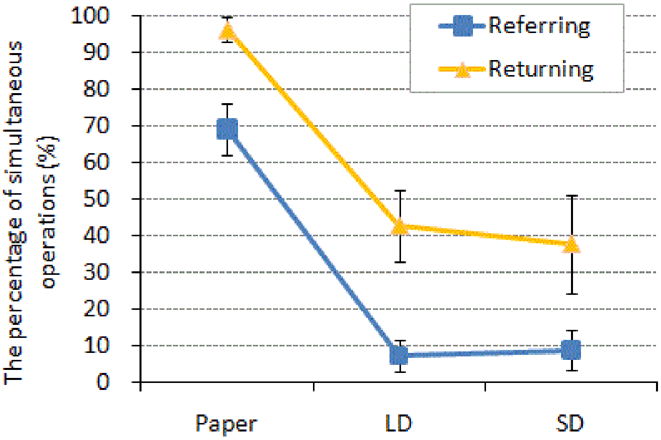


Fig. 4. The percentage of simultaneous operations

A two-way repeated measures analysis of variance was conducted to assess the reading speed according to the medium (Paper, LD, or SD) and the direction of referring (referring and turning back). Results show that the main effects of the medium [$F(2, 34)=130.43$, $p<0.001$] and the direction of referring [$F(1, 17)=98.94$, $p<0.001$]. Interaction of the two factors was not significant [$p>0.1$]. According to multiple comparison by the LSD method, more simultaneous operations existed in the Paper condition than in both the LD condition and the SD condition [$p<0.001$]. When reading from paper, readers frequently perform simultaneous operations more than 7.8 times on referring and more than 2.3 times on returning in comparison with the case of reading from an electronic medium.

Fig. 4 shows that when reading from paper, readers perform reading and turning pages simultaneously, i.e., two different actions are mutually overlapping. However, when reading from an electronic medium, readers often turn pages after finishing the reading of text, i.e. the two actions are completely separate. As Sellen and Harper [11] described, computer operations heavily rely on visual cues. To turn pages, people need to look away from text to page-turning buttons or scroll bars. It makes difficult to perform two actions of reading and turning pages simultaneously. On the other hand, when reading from paper, people can turn pages with the feeling of hands without vision. It allows to interweave two different actions. It seems that this engendered rapid reading in the Paper condition.

We next consider the reasons for reading speed differences between the LD condition and the SD condition. In the LD condition, people can view a single whole page at a glance, but in the SD condition people can view only half of the page in a display area. Therefore, we expect that people must perform more operations such as clicks or drags on a scroll bar in the SD condition than in the LD condition.

Fig. 5 shows the number of operations such as clicks on a scroll bar (Click), drags of a scroll bar (Drag), and clicks of page-turning buttons (Button) per instance of referring to notes and returning to the main text. A two-way repeated measures analysis of variance was conducted to assess the reading speed according to the medium (LD and SD) and type of operation (Click, Drag, and Button). Results show that the main effects of the medium [*F*(1, 17)=8.87, *p*<0.01] and type of operations [*F*(2, 34)=4.04, *p*<0.05] were significant. Interaction of the two factors was also significant [*F*(2, 34)=5.75, *p*<0.01]. We tested the simple main effects of the medium for each type of operations. The simple main effect was significant only for Drag [*F*(1, 17)=19.25, *p*<0.001]. Additionally, we observed a tendency by which the number of button operations in the LD condition was greater than that of the SD condition [*F*(1, 17)=3.70, *p*<0.1].

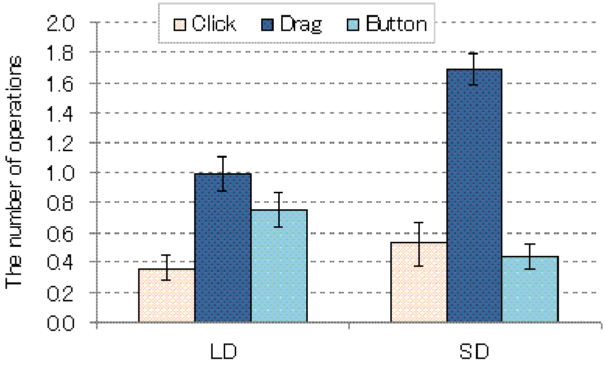


Fig. 5. Number of operations for page-turning

Using the large display, people frequently used page-turning buttons that enabled page-turning with a single click. However, with the small display, people frequently dragged a scroll bar, which seemed to take much time. Apparently, the difference of time required for referring to notes and turning back to main text between the LD condition and the SD condition is attributable to the difference of window operations.

5 Conclusion

We quantitatively compared subjects' reading performance using three media: paper, a large electronic display that presents a whole page, and a small electronic display that presents half of a page. The reading speed using paper was equal to those of both displays if the reading included no page turning. However, in the case of reading with frequent page turning, the reading speed for paper was 6.8% faster than for the large display and 11.4% faster than for the small display. A recognition test of important words of documents, we found no difference among media. Results showed that people were able to achieve the same level of comprehension as that achieved with computer displays for a short time.

To determine why reading from paper is faster than computer displays, we analyzed the process of reading in detail. Results show that subjects were able to refer to notes and return to the main text efficiently when reading from paper. Furthermore, when reading from paper, subjects frequently performed reading and page-turning simultaneously: they interwove two actions while reading. We think that this brought smooth fast reading in paper.

The results presented in this paper indicate that, when reading documents with frequent page turning, reading time could be reduced up to 11.4% without sacrificing comprehension if a person were able to select the appropriate medium. Examples of reading during which readers frequently move back and forth between pages are the following:

- reading of books with numerous notes,
- reading of documents with many references (e.g., academic paper and reports),
- reading of documents for which figures and tables are listed in the end of documents (e.g. patent descriptions), and
- proofreading while checking terminology.

We think that it is valuable to reconsider the use of paper in such instances of reading.

We are currently conducting a series of experiments to compare the performance of paper and electronic media. We intend to investigate reading of other types such as cross-reference reading using multiple documents, skimming, and reading with frequent annotation, as future research.

References

1. Duchnicky, J.L., Kolers, P.A.: Readability of text scrolled on visual display terminals as a function of window size. Human Factors 25(6), 683–692 (1983)
2. Wright, P., Lickorish, A.: Proof-reading texts on screen and paper. Behaviour and Information Technology 2(3), 227–235 (1983)

3. Mills, C.B., Weldon, L.J.: Reading text from computer screens. ACM Computing Surveys 19(4), 329–357 (1987)
4. Gould, J.D., Alfaro, L., Barnes, V., Finn, R., Grischkowsky, N., Minuto, A.: Reading is slower from CRT displays than from paper: Attempts to isolate a single-variable explanation. Human Factors 29(3), 269–299 (1987)
5. Gould, J.D., Alfaro, L., Barnes, V., Finn, R., Haupt, B., Minuto, A.: Reading from CRT displays can be as fast as reading from paper. Human Factors 29(5), 497–517 (1987)
6. Hansen, W.J., Haas, C.: Reading and writing with computers: A framework for explaining differences in performance. Communications of the ACM 31(9), 1080–1089 (1988)
7. Richardson, J., Dillon, A., McKnight, C.: The effect of window size on reading and manipulating electronic text. In: Megaw, E. (ed.) Contemporary ergonomics, pp. 474–479. Taylor & Francis, London (1989)
8. Dillon, A., Richardson, J., McKnight, C.: The effect of display size and text splitting on reading lengthy text from screen. Behaviour and Information Technology 19(3), 215–217 (1990)
9. Muter, P., Maurutto, P.: Reading and skimming from computer screens and books: The paperless office revisited? Behaviour and Information Technology 10(4), 257–266 (1991)
10. Dillon, A.: Reading from paper versus screens: A critical review of the empirical literature. Ergonomics 35(10), 1297–1326 (1992)
11. Sellen, A.J., Harper, R.J.: The myth of the paperless office. MIT Press, Cambridge (2001)
12. O'Hara, K., Sellen, A.J.: A comparison of reading paper and on-line documents. In: Proc. CHI 1997, pp. 335–342 (1997)
13. O'Hara, K.P., Taylor, A., Newman, W., Sellen, A.J.: Understanding the materiality of writing from multiple sources. International Journal of Human- Computer Studies 54(4), 269–305 (2002)
14. Schilit, B.N., Golovchinsky, G., Price, M.N.: Beyond paper: Supporting active reading with free form digital ink annotations. In: Proc. CHI 1998, pp. 249–256 (1998)
15. Chen, N., Guimbretiere, F., Agrawala, M., Lewis, C.: Enhancing document navigation tasks with a dual-display electronic reader. In: Proc. UIST 2007 (1997)
16. Holman, D., Vertegaal, R., Altosaar, M., Troje, N., Johns, D.: PaperWindows: Interaction techniques for digital paper. In: Proc. CHI 2005, pp. 591–599 (2005)
17. Piolat, A., Roussey, J.Y., Thunin, O.: Effects of screen presentation on text reading and revising. International Journal of Human-Computer Studies 47(4), 565–589 (1997)
18. Dyson, M.C., Haselgrove, M.: The influence of reading speed and line length on the effectiveness of reading from screen. International Journal of Human-Computer Studies 54(4), 585–612 (2001)
19. Sellen, A., Harper, R.J.: Paper as an analytic resource for the design of new technologies. In: Proc. CHI 1997, pp. 319–326 (1997)
20. Adler, A., Gujar, A., Harrison, B., O'Hara, K., Sellen, A.J.: A diary study of work-related reading: Design implications for digital reading devices. In: Proc. CHI 1998, pp. 241–248 (1998)
21. Sovik, N., Arntzen, O., Samuelstuen, M.: Eye- movement parameters and reading speed: A study of oral and silent reading performances of twelve-year-old children. Reading and Writing 13, 237–255 (2000)

Viability of Mobile Devices for Training Purposes

Shehan Sirigampola, Steven Zielinski, Glenn A. Martin, Jason Daly,
and Jaime Flores

Institute of Simulation and Training,
3100 Technology Parkway, Orlando FL 32826
{ssirigam,szielins,martin,jdaly,jflores}@ist.ucf.edu

Abstract. Mobile devices offer an advantageous platform on which to perform training simulations. However, they create new issues that are unique to mobile device development. First we will explore several reasons for using mobile devices for training simulators, instead of desktop or legacy-based systems. Once we have established the requirements of mobile simulators, we will discern some of the differences that arise between major mobile platforms in each development area. Finally we propose a solution that will address these differences and aid developers in creating cross-platform mobile simulations.

1 Introduction

Training simulators have been an integral part of military, commercial, and general practice towards the goal of improvement. The goal is that by doing similar activities in a similar environment, this would translate to an end result of perfection. Based on this concept people have tried to mimic the environment while minimizing the cost.

In the beginning available materials were used to create an artificial environment that had similar qualities to the desired training environment. With the introduction of computers we have improved to using graphical worlds to model the artificial environment. They were at first large legacy systems with bulky hardware and peripherals. These systems were difficult to transport and costly to build and maintain. Desktop game-based systems were then introduced, which sacrificed some realism initially in order to reduce cost and bulkiness.

Now with the introduction of mobile devices, such as smartphones and tablets, simulation-based training can evolve further. These devices offer even more portability as the simulations can now be carried to training sites with ease. Mobile systems also cost less than most desktop simulation systems. Thus it is advantageous to find a solution that provides an easy transition from desktop to mobile platforms.

2 Mobile Simulation

Simulations have a plethora of uses, and mobile simulations may have even more. They can be used for training as well as the display and analysis of scenarios. Our focus is in two areas: training simulations and training support applications (initially, after-action review). Training simulations encompass a wide range of domains (combat, medical or otherwise). The only limit to the type of training is that the

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 102–109, 2011.

platform must be able to robustly simulate the scenario. After-action review allow for in-depth analysis of training exercises. Examples of this type of system are DIVAARS and DART [1].

2.1 Criteria

Training applications have certain criteria that must be maintained on mobile platforms. We shall now explore these criteria in detail.

Represent Diverse Environments. As stated earlier, we want to simulate a wide variety of scenarios. For example simulations may be required to simulate outdoor and indoor environments. We also may need to make drastically different simulation scenarios such as the inside of an artery. Thus the simulations must be able to deliver output in a variety of methods (e.g.: 2-dimensional, 3-dimensional, and tactile).

Compatible. In order to keep up with today's mobile technology the simulations must run on a variety of platforms. Specifically the simulations should cover the major smartphone platforms. In this paper the Android and iOS platforms receive the most focus, as they are two of the most popular platforms in the market today. However there are other platforms that need to be considered, such as Symbian, Meego and Windows Mobile. The simulations should also have at least one method with which to interact with one another, such as Wi-Fi or Bluetooth, in order to support collaborative exercises.

Performance. The simulations must run within acceptable bounds of mobile platform performance. These platforms usually have less performance than desktop and legacy systems. Thus we have to make sure to optimize many of the aspects of the simulation. For example, 3-dimensional graphics is an area that uses a lot the resources of the platform. We must make sure that the process of drawing the virtual scene is optimized so that unnecessary parts of the environment are not drawn and that the number of calculations is minimized.

Scalable. Another criterion for the mobile simulations is to be able to link together in order to create a distributed simulation. The basics of this requirement were covered in the compatibility requirement, but there are more sophisticated issues that arise. Such an issue is the virtual environment represented by the simulation needs to scale in the amount of operations it needs to performed as additional users join the exercise. Otherwise, if there are few users in the environment, then the simulation must scale down in order to preserve resources, such as battery life.

Portable. Portability is one of the major requirements of a mobile simulation. This means that the simulation should not contain bulky peripherals, or if the mobile platform itself is too bulky, then it should not be used. Of course the mobile simulation solution must still have the ability to perform adequately despite these portable bounds.

Ease of Use. Most importantly, a mobile simulation must be easy to use and should not overtax the operators. This means that they must contain user friendly interaction methods and interfaces. This is particularly important as mobile platforms contain small screen sizes; therefore they must not flood the screen with too much information.

Solution. The mobile simulation solution must enable programmers who use it to create simulations that abide by these requirements. It must provide high-level functionality that will allow relatively easy creation of efficient simulations without taking away too much control from the programmer, so that they are able to represent their environment accurately.

3 Challenges, Differences and Solutions

Now that the criteria for a mobile simulation have been established, let us approach the implementation of the solution. The greatest barrier to this implementation is that mobile platforms are vastly different in their implementations and behaviors. There are also challenges unique to mobile platforms that have not been encountered in other simulation platforms. Each area of mobile simulation development brings its own issues and requires its own solution. This section explores those areas, their issues and proposes solutions to those issues.

3.1 User Interface

The major difference that needs to be overcome between multiple phones is the development of different user interface elements. To begin, each mobile platform comes with its own suggested coding environment. They also give different ways to interact with their user interface elements. For example, the Android platform has the developer either use an Eclipse plugin or use XML to represent the user interface elements; the iOS platform requires the developer to use the proprietary Interface Builder that is only available on Macintosh computers.

The use of most of these can be ignored if all user elements are custom written. However, this is usually not an option in most cases due to time constraints. Therefore, this requires the use of both user interface approaches for the design of any application. This then leads to a desire for the developer to write code to control the user interface without recoding the program for each mobile platform. In order to accomplish this, the user of the solution must be able to use the given interface designers and then use a library to write the code for their application.

Next, for each individual interface element we need a way to acquire the element from the environment. Again, there are major differences here. The Android platform uses an enumeration and specific Java methods to acquire the user interface element. iOS uses connections that were built through the Interface Builder. This leads to problems when trying to access the user interface elements in code.

A solution to this issue is to use message passing in order to acquire the information necessary. For Java this is simple and simply requires a String and access to the Java object from which the enumerations are stated. However, complications arise in the iOS platform. We have to search through the XML document provided by the Interface Builder to locate the element.

Once the interface elements have been created the user needs to be able to change and manipulate them during runtime. This is normally done by storing each element to a different variable of a given class and accessing the methods of the class. Other instances like user interaction require the implementation of specific methods. On iOS this requires the use of the Interface Builder in order to connect the specified user interface element with the proper method. In Android you attach the developer's method with a method call that gives the developer's method as a parameter.

Our proposed solution stores the different user interface elements into their most generalized type. This allows it to group all user interface elements into large storage structures. So when an element is needed all the user receives is the location in our storage structure that has the correct element. Next, when calling a method we are able to cast the element to the class that contains the specified method and then call the method with the user given input. Should the user need to write a method to handle input we propose using an instance of an abstract class.

3.2 Application Structure

When programming for a mobile platform the first lesson taught is the application lifecycle. The lifecycle controls the flow of any program from start to finish.

The application lifecycle between the two phones is radically different as shown in figures 1 and 2 [2][3]. iOS relies heavily on the event loop and a function similar to Android's. However, Android uses many different points throughout the applications execution in order to allow the programmer more control over the situation.

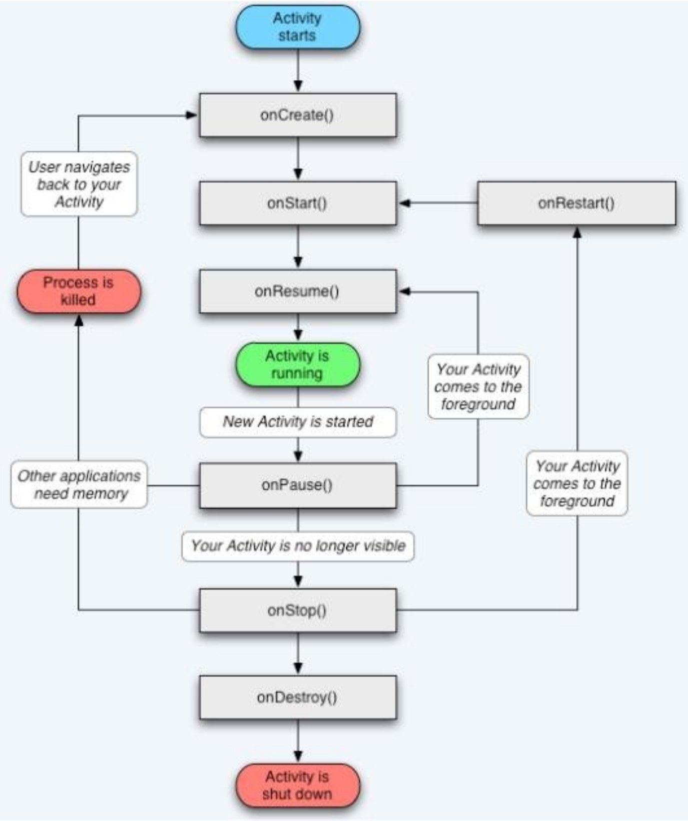

Activity starts
onCreate()
User navigates back to your Activity
onStart()
onRestart()
Process is killed
onResume()
Activity is running
Your Activity comes to the foreground
New Activity is started
Other applications need memory
onPause()
Your Activity comes to the foreground
Your Activity is no longer visible
onStop()
onDestroy()
Activity is shut down

Fig. 1. The Android lifecycle [2]

Our proposed solution to this lies in trying to separate the two lifecycles and abstracting this implementation from the user. This is done by taking all of the Java calls and transferring them to specific methods in our library. From there we allow the user to access specific information when a specific call is made by the implementation. The simpler structure of iOS allows for a simpler solution; where only a few methods are implemented that, when called, send that call to the user to determine what to do through the same predefined abstract class.

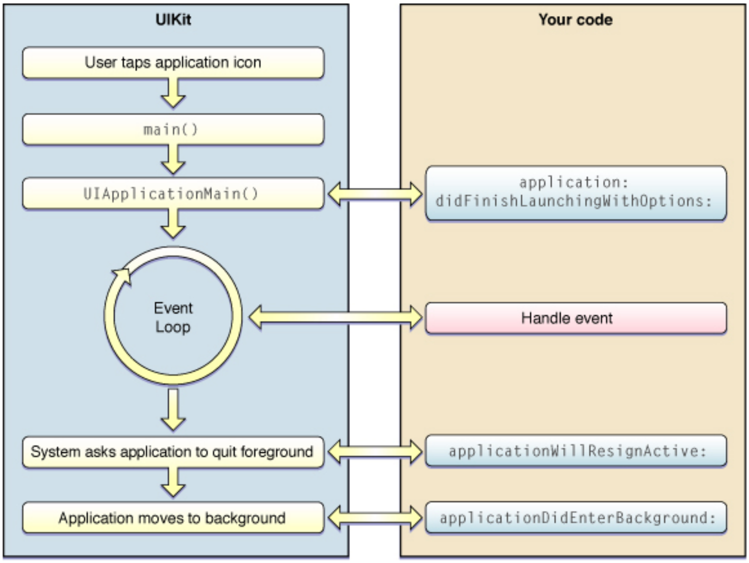

UIKit
Your code
User taps application icon
main()
UIApplicationMain()
application:
didFinishLaunchingWithOptions:
Event
Loop
Handle event
System asks application to quit foreground
applicationWillResignActive:
Application moves to background
applicationDidEnterBackground:

Fig. 2. The iOS Lifecycle [3]

3.3 Programming Languages

A further complication is that each mobile platform has its own programming language that is used. For example iOS uses Objective-C, Android uses Java, and Windows phone uses C#. Due to this every program written for one platform would need to be rewritten to match the respective programming language of the platform. The differences between these are small enough that the changes would not require a complete rework of algorithms and design, but large enough that most applications would a significant effort to translate the code into each platform.

Our solution is to use a language that is supported by all the mobile devices. C++ serves as a good candidate for this role. Although the standard development kit for Android did not initially support C++, they have since released the Native Development Kit, which now allows C++ code to run on the Android. The iOS platform allows C++ code from the very start. However, certain methods and classes had to be written in Objective-C, since C++ cannot be directly accessed from Objective-C code. Therefore, our proposed solution contains a given set of classes that will interface with the Objective-C core of the program at the start, and then take over the remaining functionality of the program.

3.4 Three-Dimensional Graphics

Three-dimensional graphics on mobile platforms have evolved quite a bit over the past several years. This progress has been greatly aided by the introduction of OpenGL ES graphics, which was specifically developed for embedded devices. It has been implemented in the iOS and Android platforms as well as on some Symbian and Maemo systems. Thus our 3-d graphics solution will focus on the use of OpenGL ES as a base. Specifically the 2.x versions of OpenGL ES introduce the use of programmable shaders in place of a fixed-function pipeline. This development allows for greater flexibility in graphics programming as programmers can now specify how they want objects to be drawn onto the screen.

OpenGL ES delivers the basic functionality required to program graphical scenes, but it lacks many high-level functions that are required for successful 3-d simulation programming. In order to meet our simulation criteria the 3-d implementation should meet performance requisites and allow for fine-grained control so that users may create a variety of environments to suit their needs. Thus, we add another layer of abstraction that delivers both of these requirements.

Several libraries and development kits have been created on top of OpenGL ES in order to aid mobile graphical programming. Examples of these solutions are Shiva 3D, Unity, Airplay SDK and OpenSceneGraph [4][5][6][7]. Shiva 3D and Unity offer an easy-to-use interface and contains many features, but require the use of their proprietary editor in order to create mobile applications. This creates a loss of fine-grained control as the user does not have full control of the implementation of the application. This extends outside just the 3-d graphics area and restricts control over the user interface, connectivity and other areas.

The Airplay Software Development Kit (SDK) does not require the use of any proprietary editor and thus offers great control. As the name implies, it allows the user to create their programs with minimal hindrance. However, this SDK lacks the high level functionality that is required. For instance, the SDK lacks a sophisticated method of organizing a 3-d scene. An organized scene allows for better optimizations by reducing the amount of state changes that are required to draw the visible scene. These types of optimizations allow for a much smoother performance, especially on the mobile platforms with limited resources.

The OpenSceneGraph (OSG) library contains both rich optimization features and great control. However, the library has not been fully implemented on mobile platforms. Currently developers are creating a version for iOS, but the library has not been implemented on Android or any other platform as of the writing of this paper. A significant effort would be required to adapt the library to these other platforms. The use of OSG also raises the question of whether its form of scene organization is well suited for mobile development. OSG uses a scene graph, which is a hierarchical representation based on objects with children nodes representing the sub-objects of that object. This has been well-suited for desktop simulation programs and legacy simulators, but has yet to be tested thoroughly on mobile platforms.

Thus, we are pursuing solutions that either implement OSG on several mobile platforms or create a new abstraction layer that embodies the requirements. The latter would require an even more significant effort, but would yield the most desirable results, since the library can be optimized specifically towards simulation programming on mobile platforms.

3.5 Connectivity

Connectivity serves an important purpose as it allows the mobile platforms to communicate with each other and create distributed simulation environments. Two main methods of connectivity are Wi-Fi and Bluetooth. These allow for a wide range of connectivity scenarios of different range and scale. Each mobile platform contains its own different implementation of connectivity with its own security protocols. For example, the iOS platform only allows access to specific Bluetooth profiles, while other platforms allow its usage with the user's permission.

Since such diversity exists between the phones, the solution must provide a large layer of abstraction that veils the implementation details and the specific means of the connection. The interface for this layer should also not distinguish between the type of connection (Wi-Fi, Bluetooth), so as to allow the simulation programmer to specify the connection type yet, not tax the programmer with details of platform-specific security and implementation.

4 Conclusion

Even though mobile simulation platforms may have limited resources, their use would provide noticeable benefits by lowering cost as well as increasing portability and ease. However, in order to take full advantage of the great diversity of mobile platforms, one would have to create several different versions of a simulation, which would require significant time and effort. This is caused by the different implementations of graphics, user interface elements, and connectivity. We proposed a solution that should mitigate these differences through layers of abstraction. Thus our proposed solution should considerably aid in writing mobile simulations by enabling users to design their specific virtual environments without having to worry about technical and implementation details that are specific to each mobile platform.

Using our specified design we plan to continue to implement our proposed solution. Once finished, we will test to confirm the reliability and efficiency of the sub solutions until they achieve the necessary criteria. Finally they must be packaged together to create a complete solution that will provide a suite for mobile simulation programmers.

References

1. Knerr, B.W., Lampton, D.R., Martin, G.A., Washburn, D.A., Cope, D.: Developing an After Action Review System for Virtual Dismounted Infantry Simulations. In: Interservice/Industry Training, Systems and Education Conference (I/ITSEC 2002), Orlando (2002)
2. Android Application Lifecycle, http://www.androidjavadoc.com/1.0_r1_src/android/app/Activity.html

3. Iphone Application Lifecycle, `http://developer.apple.com/library/ios/#documentation/iphone/conceptual/iphoneosprogrammingguide/CoreApplication/CoreApplication.html`
4. Shiva 3D, `http://www.stonetrip.com`
5. Unity, `http://unity3d.com`
6. Airplay Software Development Kit, `http://www.airplaysdk.com`
7. OpenSceneGraph, `http://www.openscenegraph.org`

Display System for Advertising Image with Scent and Psychological Effect

Keisuke Tomono[1], Hiroki Wakatsuki[2], Shigeki Kumazawa[2], and Akira Tomono[2]

[1] Department of Chemistry/Engineering Science, Wartburg College
100 Wartburg Blud, Waverly Iowa 50677-0903, USA
[2] Department of Information Media Technology, School of Information and
Telecommunication Engineering, Tokai University
2-3-23 Takanawa, Minato-ku, Tokyo 108-8619 Japan
tomono@keyaki.cc.u-tokai.ac.jp

Abstract. We propose a new method in which scents are ejected through the display screen in the direction of a viewer in order to enhance the reality of the visual images. A thin LED display panel filled with tiny pores was made for this experiment, and an air control system using a blower was placed behind the screen. We proved that the direction of airflow was controlled and scents properly travelled through the pores to the front side of the screen. Moreover, the effectiveness to an advertising field of this system was estimated by simulating an actual situation in which various advertisements are around using the immersive VR System. The subjects' eye movements and impressions, when they look at the scented advertisements while walking, were analyzed.

Keywords: Display, Scent, Digital signage, Multi-Media, Advertisement.

1 Introduction

Because the technology of making a large and thin display screen has advanced in recent years, as for the graphic display device, the restriction of the installation location has decreased. It is possible to hang or integrate it with the wall. When this technology is used as digital signage, the advertisement has large flexibility, compared with the signboard that uses the photograph [1]. However, the digital signage should improve its cost-effectiveness because it is expensive. For instance, the ability to catch the eye and urge a passer-by to stop in front of the signboard is necessary.

The scent emitting systems to attract passers-by have been greatly developed, and these systems have started to be applied to various advertisements in places such as supermarkets and restaurants [2], [3]. From a merchandising prospect, these systems are expected to attract the passer-by to the place, which sells the product that relates to the scent and improve the desire to purchase it. Therefore, a highly effective advertising advantage can be expected by the addition of the scent corresponding to the product advertised on the digital signage. However, a conventional olfactory display only provides scents not accompanied by visuals. Therefore, when the product are introduced using the image, it is necessary to allocate an olfactory display near the graphic display device. Moreover, a conventional system has the problem of taking up too much space in the installation.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 110–119, 2011.

In this paper, in order to solve the problems above, a new display system called the KANSEI Multi-Media Display (KMMD) that can present the visuals and scents with one device is proposed [4], [5]. To verify this concept, the display device that the air passed from the hole installed between the pixels of the thin display panel was made for trial purposes, and the characteristics of the air flow and scent presentation to the viewer were evaluated. Next, a psychological effect of the advertisements with scents, based on the concept of the scented digital signage, was evaluated by an experiment using subjects. Since it is difficult to evaluate it in the environment, where the digital signage is actually set up, it was evaluated in a virtual town that was created by using an immersive Virtual Reality (VR) system.

2 Related Works

The large display system that uses an LED for the pixel is often used in digital signage. In case that the screen size is 3×2m, and the resolution is 752×528 pixels to which the television image of NTSC standard can be displayed, the pixel pitch is about 4mm [6]. Regarding the display system using an Organic LED (OLED) as pixels [7], the large size one which was 3456×2471 mm (width × height) and had 1152×640 pixels with 3mm of pixel pitch, which has a function of High Definition Television (HDTV) was introduced from Mitsubishi Electric Corporation in CEATEC Japan in 2009. Thus, it is thought that a practical enough resolution is obtained for the pixel-pitch of 3~4mm in the digital signage of the width of about 3m usually seen in a town. However, as for the display device, the research that discharges the air or scent through the hole installed on the screen doesn't exist.

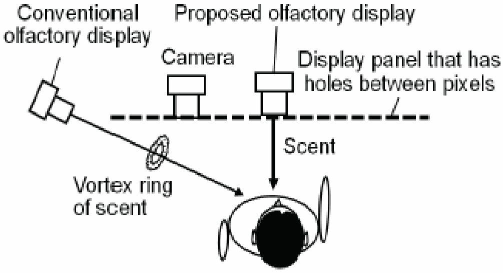

Conventional olfactory display
Proposed olfactory display
Camera
Display panel that has holes between pixels
Scent
Vortex ring of scent

Fig. 1. Concept of the KMMD

As for the unwearable olfactory display, the air cannon type was developed [8]. This system is able to emit air-rings through air cannons from a short distance away from a viewer. However, there is still not much research regarding the emission of scents through holes in the screen.

3 KANSEI Multi-Media Display (KMMD)

In order to synchronously present a visual and a scent, the most efficient way to advertise is to emit a scent to a viewer's nose. Therefore, a scent emitting device using an air cannon has been developed [[9]. By detecting the position of one's nose with an image processor and discharging an air ring with a scent to their direction, the scent can be presented in a particular place at a particular time. However, as Figure 1 shows, because it is necessary to emit a scent from the side of the screen, if the screen is large, it is difficult to carry an air ring with a scent for a long distance.

On the other hand, the KMMD has the following feature. A thin display panel that has many tiny holes between pixels is used. A video camera and scents-emitting device are placed behind the display panel, and the scents are emitted through the holes in the screen to a person who is detected by the monitor camera. Since the distance between the olfactory display system and the person is relatively short, the system has the advantage of presenting a scent from an area near the displayed object toward the viewer with certainty. Therefore, the presence is obtained as if the scent drifted from the product displayed on the screen.

3.1 Prototype of LED Display That Air Passes through

If the hole installed on the screen is easily recognized, one's mental image associated with the objects on the screen is destroyed, and the quality of the image might be decreased as a result. Then in order to understand an appropriate size of the holes which can be installed in the display panel with 3 to 4 mm of pixel pitch, the thin type LED display panel was ordered for this experiment. The pixel pitch of this panel corresponds to the resolution of the digital signage with image quality of standard television level described in Chapter 2. This prototype LED display is used to clarify the mechanism of emitting gas to a particular direction. Since the whole experiment is just to confirm our principle, a compact size of a display was designed to present simple diagrams and for a user to be able to clearly read scripts such as Kanji, the alphabet, and numbers.

We used a 1.6 mm thickness of a FR-4 epoxy-glass which is 300mmW×200mmH. 2.4 mmφ diameter holes were created at intervals of 4 mm in the spaces between the pixels in the display area. The LEDs were mounted at 4 mm intervals. The area of the holes accounts for 28 percent of the display screen. On the top of the board, a compact CCD camera was mounted in the 2.4 mmφ hole. Scents were ejected toward the viewer who was monitored by the camera.

3.2 Display Characteristic

Figure 2 shows a prototypical display which was made for this experiment. Picture (A) is the face and screen side. Picture (B) is the back. The color of the board on the face is a pale gray color. The brightness of the LEDs is controlled in 10 stages. On the screen of the display, sentences in Japanese or English are described and run from right to left, and the size of each character that composes the sentences is 50 to 60 mm. In order to evaluate the visibility of the words, a questionnaire was given to each subject and their answers were totalized to a sum. By changing the brightness of the

LED pixels, the subjects were asked about the visibility of the objects on the screen. They had to choose between one of five choices ranging from “not really able to read,” to “clearly able to read” and “comfortably able to read.” The average from the seven subjects is shown in Figure 3. These people were also asked if they recognized the holes on the screen and, if so, the subjects were then asked if these holes interfered with the readability of the screen. A summary is shown in Figure 4.

Before the experiment began, our main concern was whether or not the subject could recognize the holes on the screen, because the LED pixels had the potential to illuminate the holes around the LED pixels themselves. However, the holes were not always recognized by the viewers during the experiment in the dark room. It is understood that this was because the LED pixels were mostly illuminated from the front and not from the sides. Another reason is because the human ocular system tends not to perceive objects that are near shining objects.

Since the display panel which was created for this experiment was small, complex and large visual images were not displayed on the screen. If the screen size is enlarged to a width of about 3m, it seems that the standard television quality can be achieved because the source pitch is 4mm.

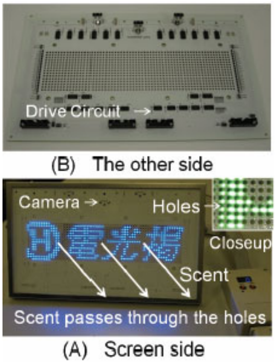


Fig. 2. Display with many holes

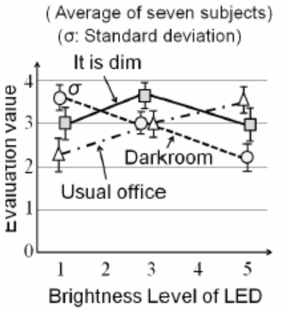


Fig. 3. Readability of displayed characters

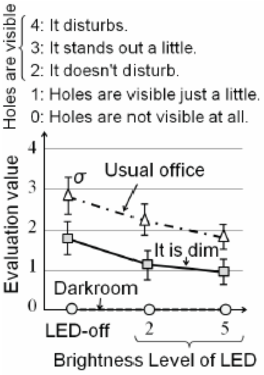


Fig. 4. Visibility of holes

4 Olfactory Display System Behind the Screen

4.1 Air Discharge Mechanism Design

KMMD uses an air control system which was placed behind the thin display with tiny holes to change the air pressure as shown in Figure 5. The control system was composed of an air compression mechanism, an air controller box, and an air duct. Compressed air was carried to the air controller box by the air duct. The air controller box was tightly attached to a specific position on the back of the screen. Therefore, the pressurized air in the box was smoothly emitted through the porous screen. By putting scents into the box, vaporized scents were discharged with the airflow. The blades were installed in the air controller. We thought that the air passed through in various directions through the holes on the screen by changing the angle of the blades because the display panel had a thin design.

4.2 Airflow Control Experiment

We experimentally verified the flow control of air. The measurement conditions and the summary are shown in Figure 6. The eleven blades were adjusted to a specific angle to the display, and a distance was set between the display and each subject. The wind velocities directed horizontally to the display were analyzed using the Hot Wire Anemometer, RoHS, DT-8880. The subjects' reactions to each wind velocity were evaluated through answers to a questionnaire.

Picture (A) from Figure 6 shows the average velocity for ten seconds when adjusting the blades to 30 degrees to the display as an example, with a distance of 30 cm between the display and each subject. It is understood that the direction of the air flow changes by rotating the blades.

By changing the output of the blower, the airflow could be gradually adjusted between a gentle and heavy breeze at any distance up to 60 cm from the display. Since the strength and the direction of the airflow could be controlled, it will be possible to use it for not only presenting scents but also presenting the current of air like the wind to the skin, if a large-scale screen is made in the future.

4.3 Scent Emitting Experiment

A porous material (a cotton ball) was soaked in essential oil, such as vanilla and was placed into the air controller box where the oil was vaporized. The vaporized scents were discharged through the porous holes with either a gentle or weak breeze. As shown in Figure 6, changing the direction of the airflow made subjects feel the strongest scents in that area where the wind velocity was the strongest. This occurs because the scents were more concentrated in that area. This result could be applied to the situations of scent emissions coming from areas near the visuals on the screen.

5 Psychological Evaluation of Digital Signage with Scent

In a psychological examination of the advertisement using a scent, the eye catching effects were first examined. Next, the effects of scent presentation on one's memories were examined. In this experiment, a simulation experiment was also conducted using the immersive VR System, HoloStage$^{\text{TM}}$ made by the Christie company, because there were a lot of restrictions to the experiment in an actual passageway.

5.1 Eye Catching Effects of Advertising Image with Scent

Experimental Method. Figure 7(a) shows the virtual town used for the experiment. The scale of the VR space is equal to a real space, and the length of the passage is about 20m. Five digital signages were placed in a line on the left wall, and objects of the person type and the ornament, etc. were set up at the right of the passage. The digital signages for a cafe, roast meat, fruit, pizza, and flowers were queued up from this side sequentially on the left wall. The contents for each digital signage used animation. The digital signages with the scent were evaluated in two kinds of patterns. In the experiment for pattern A , when a roast meat shop's signboard or a flower shop's signboard appears, the roast meat scent or the rose scent were presented respectively. In the experiment for pattern B, when a cafe's signboard or a fruits shop's signboard appear, the coffee or apple scent were presented respectively.

Figure 7(b) shows the appearance of the experiment. The procedure of the experiment is as follows. The subject advanced in the passage by the feeling to take the moving sidewalk, because the experimenter drew the VR space forward by the walking rate. The time to pass over the passage was about 40 seconds. The eye movement data of this time were accumulated, and the gaze of the objects were analyzed. The scent were presented not to be noticed from the back of the subject for about 5-7 seconds with a scent spreader when the subject approached the digital signage of the target. The subject obtained the sense of which the scent drifts from the passage while walking. The gaze objects when the scent was presented or not were compared. The subjects were questioned about their impression after the experiment ended.

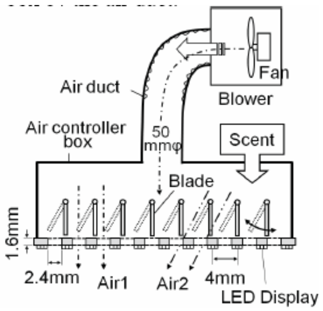


Fig. 5. Air and scent discharge mechanism

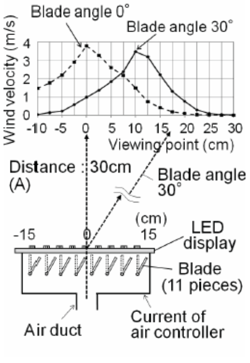
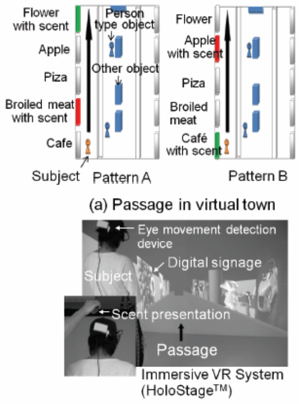


Fig. 6. Experimental results that control the air flow direction



(a) Passage in virtual town

(b) Experimental environment

Fig. 7. An experiment of the eye catching property

Results. Figure 8 shows one example of the gaze detection result. The vertical axis are the objects that exists in the passage, the belt of a horizontal axis is time to gaze at each object. It was understood that the subjects gazed at the signboard with a scent at a high frequency compared to without the scent in both pattern A and B. Figure 9 shows the relation between the kind of the signboard and the average gaze time of ten of the subjects. The signboards that presented the scent were gazed at for a long time. Also, all subjects answered that he or she looked for the signboard when the scent was felt, and were interested in it. Thus, it was understood that the sense of smell was stimulated by presenting the scent and the digital signages with scents were considered strongly.

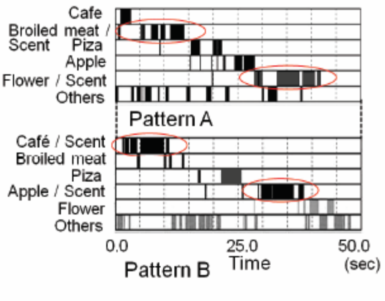


Fig. 8. An example of the change of the gazed target

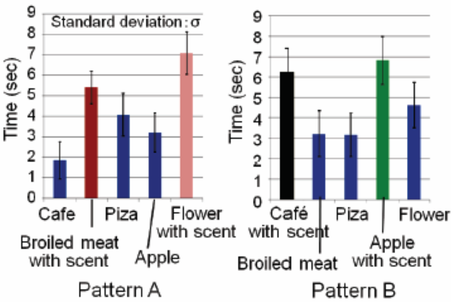


Fig. 9. Average gaze time of 10 subjects

5.2 Memory Promoting Effects of Image with Scent

A lot of clues for recollection are enumerated as for storage by combining different symbols. Then, the advantages to the memory of the image with the scent was investigated. Moreover, to explain the results objectively, the biological reactions when memorizing and recollecting them were examined.

Experimental Method

1. Method of memorization

One image was presented from 15 kinds of flower images on the screen of 120 inches one after another. The tasks of generating a mental image to the object and memorizing it were given to the subjects. One image was presented for 15 seconds. When the image was switched, a 30 second rest was taken to stabilize their feelings. Here, the scents added three images among 15. The scent was presented under one's nose by strength to which the kind was understood.

Figure 10 shows the appearance of the experiment. The sensors of Near Infra-Red Spectroscopy (NIRS) that measured the changes in the brain hemoglobin concentration were attached to the right and left sides of the subjects' frontal lobes. This NIRS was made by Hamamatsu Photonics (Multi-Fiber Adapter (MFA) for NIRO-200). Also, a biological sensor, Polymate 2 (TEAC Corporation), was used to analyze the skin conductance. A pair of active biological electrodes to analyze skin conductance were attached to two fingers of each subject's left hand.

2. Method of testing recollection

After memorizing 15 images, the rest was taken for about five minutes, then, the test images were presented one after another from 25 images that contained all of the 15 images used for memory, and when the same kind of image as the memorized image was recognized, the subject was instructed to push the button. Here, the subject was directed to push the button when it was possible to confirm it firmly, because it was not a test of speed. Since the purpose was to examine the relationship between the depth of the processing level of memorizing the image with the scent and the result of recollection, in the recognizing task, the scent was not used. Also, the test image of the composition was different from the image that had been used when the subject was memorizing the images even if they were the same kind of flower image.

Each image was presented for ten seconds, and when the image was switched, they took a 30 second break. The subjects consisted of ten seniors. The correct answer rate and reaction time from the image presentation of the answer were examined.

Results. Figure 11 (left side) shows the reaction time comparison from the image presentation to the answer on the presence of the scent condition. It is understood that the judgment is early in the image with scent. Also, Figure 11 (right side) shows the correct answer rate in the recollection test. The image with the scent is recollected accurately. In the questionnaire after the examination, the following answers were obtained as an impression. "As for the image with the scent, the scent strongly leaves an impression.", "When memorizing it, the scent that had been smelt before was recalled, and the scene that related to the scent was useful for the memory." It is understood that the image with a scent strongly remains in the impression, and is effective for recalling the objects. We may assume that subjects perceive the object more consciously by recalling their past experience.

Figure 12 shows an example of the measurement result of the haemoglobin in the brain. In comparison with no scents, the concentration of the oxidized-hemoglobin increased while the deoxidized-hemoglobin decreased after being presented by scents. The results of presuming the revitalization of the brain on the scent presentation from the measurement were obtained from two or more subjects.

Figure 13 shows an example of measuring the skin conductance. When the image was presented, the skin conductances were the same as normal circumstances or rose more than normal circumstances in all ten subjects, regardless of the presence of the scent. When the rising values of each skin conductance had been compared, the image with the scent rose more than when there was no scent for eight subjects in ten subjects. It can be considered that the images with a scent cause the tension and the excitement easily as the results.

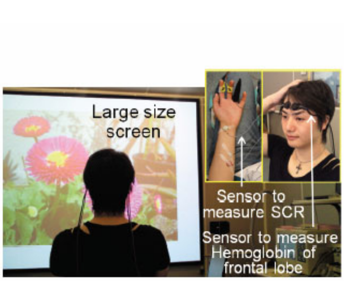


Fig. 10. Experimental environment to evaluate the psychological effect

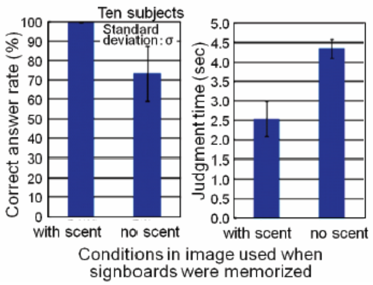


Fig. 11. Experimental results of recalling signboard presented ahead

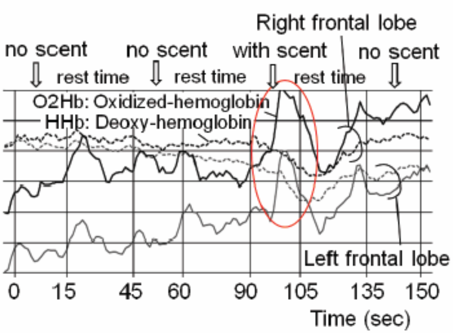


Fig. 12. Experimental results that show brain activity by scent presentation

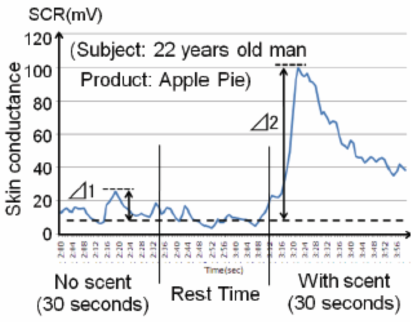


Fig. 13. Skin conductance changes by scent presentation

6 Conclusion

We made the KMMD that emit the scent through the screen for trial purposes, and examined the possibility of the scented digital signage by using the simulation. The conclusions in this experiment were as follows:

1. Three-colored LED pixels were separated by 4 mm spaces on a circuit board with a thickness of about 1.6 mm, and holes which were 2.4 mmφ in diameter were made in the spaces. The holes are considered not to degrade the quality of the visuals.
2. By tightly attaching an air controller box to the back of the LED panel, the direction of the airflow was controlled. Also, by putting scented essential oil in the box, the system was able to more naturally present the scents to the viewer.
3. The effect of the improvement to catch the eye with the image emitting a scent was shown through the simulation experiment using the Immersive VR System.
4. Presenting visuals with corresponding scents makes the viewers generate the mental image easily, and remain in their memory easily.

Though the KMMD made for trial purposes in this study was small in size, if it is possible to make it in a large-scale size, and emit the scent from the vicinity of the displayed object, one may feel as if actual scents come from the object. In the field of advertising, which uses devices such as digital signage, the display screen is used to increase a person's desire to buy the product. We want to use the KMMD also for other various fields such as virtual reality simulations and sensory games in the future.

Acknowledgements. Finally, we appreciate Mr. Hideo Kawashima and his company who cooperated in creating the LED display panel prototype.

References

1. Muramoto, K.: The Trend of Digital Signage. The Journal of The Institutes of Image Information and Television Engineers 65(2), 119–120 (2011)
2. Nakamoto, T., et al.: Olfactory display (Multimedia tool for presenting scents). Fragrance Journal (2008)
3. Tonoike, M., et al.: Information and Communication Technology of Olfaction. Fragrance Journal (2007)
4. Tomono, A., Tomono, K.: Display. PCT/JP2008/051387 (WO2008/093721)
5. Tomono, K., Tomono, A.: Display or Lighting System. Japanese patent application, JP2010-196436 (2007)
6. Musgrave, G.: Very Large-Screen Video Displays. Conceptron Associates, 2–6 (2001), http://www.conceptron.com/articles/article_index.html
7. Shinar, J.: Organic Light-Emitting Devices: A Survey. Springer/AIP-Press, Berlin (2004)
8. Yanagida, Y., Kawato, S., Noma, H., Tomono, A., Tetsutani, N.: Projection-based Olfactory Display with Nose-tracking. Proceedings of IEEE Virtual Reality, 43–50 (2004)
9. Tomono, A., Tomono, T., Fukiura, T., Yamaguchi, H.: Olfaction characteristics improvement of projection-based olfactory display. The Journal of the Institute of Image Electronics Engineers of Japan 37(4), 444–451 (2008)

Subitizing-Counting Analogue Observed in a Fast Multi-tapping Task

Hiro-Fumi Yanai, Kyouhei Kurosawa, and Kousuke Takahashi

Department of Media and Telecommunications, Ibaraki University,
Hitachi, Ibaraki 316-8511, Japan
hfy@ieee.org

Abstract. A widely-used method for entering texts with mobile phones is the multi-tapping one. To understand the internal processes of humans while doing multi-tapping, we designed an experiment with time pressure, that is, to multi-tap a key for the prescribed number of times as fast as possible. We observed three types in time series data for inter-tap intervals: Type I (Plan and Do), Type II (Do and Adjust), and Type III (Mixture of Type I and II). And, for the distribution of errors in the tapping, we found a subitizing-counting analogue. That is, if the instructed number of tapping was smaller (< 4), the error rate was smaller, and if the number was larger (> 4), the error rate rose abruptly. These findings could lead to the model of human cognition and manipulation of the number, hence to the design of the usable human interface.

Keywords: subitizing, counting, numerosity, cognitive process, text entry, mobile phone.

1 Introduction

Despite the popularity of full keyboards in mobile information devices, a lot of people still use the multi-tapping methods for text entry with mobile phones. The multi-tapping is the method of entering letters by pressing keys for multiple times [1]. When you write an e-mail with your mobile phone, you need to tap the keys multiple times. When, for example, you want to enter the letter “a”, you tap the key “2” once. And when you want to enter the letter “c”, you tap the same key three times. In the case you want to enter the letter “A” (capital A), the usual way is to tap the key four times, and if “C”, tap six times. In the standard assignment of the English alphabet, the key “9” is for the letters “w”, “x”, “y”, “z”. Thus if you want to enter “Z”, you have to tap the key eight times. In the English text entry, three letters are assigned to each of the key “2” through “9” with the exception of four letters to the key “9”. Depending on the language, the assigned number of letters per key varies. In Japanese, for example, basically five letters are assigned to the keys “1” through “0” excepting three letters to the key “8” and “9”.

To simplify the argument, we designed an experiment in which the participants were instructed to tap a single key for the prescribed number of times as fast as possible. Surely it is too simple, since in reality the text entry task involves several sub-tasks such as “determining the appropriate key”, “searching the key visually”, “targeting the key”, and “tapping the key for the appropriate number of times”. The present study focuses on the last sub-task of the text entry procedure.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 120–125, 2011.

When we react to the number, i.e. recognize the number of objects presented visually, it is known that we switch between the two modes according to the number we have to recognize. For the number less than or equal to four, we use subitizing, and for the number greater than four we use counting. To put it simply, subitizing is the one-shot recognition of the number requiring tens of milliseconds per object, and counting is a sequential procedure requiring hundreds of milliseconds per object [2]. The subitizing-counting mode is not specific to visual stimuli, but observed in auditory stimuli as well [3].

A major finding of the present study is that there is subitizing-counting analogue in the results of the fast-tapping task. We found the human error distribution in fast multi-tapping, which is analogous to the subitizing-counting of numerosity recognition.

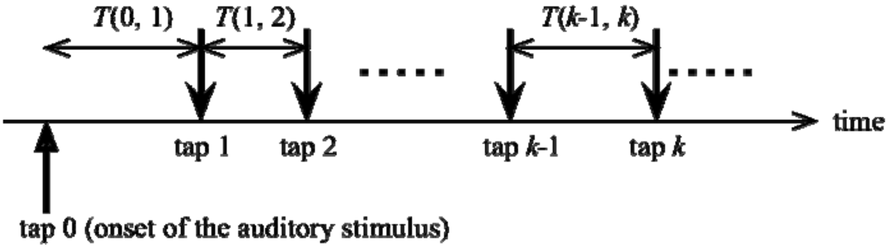


Fig. 1. Explanation of the inter-tap intervals *T*(*k*–1, *k*)

2 Experiment

Sixteen volunteers participated in the experiment. All were university students at the age of 22–24. The participants tapped a single key on the computer keyboard as fast as possible for the prescribed number of times. To minimize the other factors than just tapping, we used auditory stimuli that say “one”, “two”,..., and “nine” with the recorded human voice. And there were no visual or auditory feedback while performing the task. Stimuli were presented thirty times for each participant in the random order. Time series of every key tap were monitored by and stored in a personal computer.

The results are summarized in the following from two points of view. One is the type of time series of inter-tap intervals, and the other is the error distribution over instructed number of taps.

3 Results

The results are summarized in the following from the two points of view. One is the type of time series of inter-tap intervals, and the other is the error distribution among prescribed number of taps. First we show the inter-tap interval time series, and then show error distribution.

3.1 Time Series of Inter-tap Intervals

Inter-tap intervals are the time between successive taps *T*(0, 1), *T*(1, 2), ..., *T*(*k* –1, *k*), ..., *T*(*n*–1, *n*), where “tap 0” is the onset of the auditory stimuli (no tap) and *n* is the prescribed number (See Fig. 1). In Fig. 2, three types of responses from the participants are shown schematically. By noticing the lines connecting *T*(0, 1) and *T*(1, 2), we categorized the data into three types:

1. Type I (Plan and Do). There are longer delays in doing the first tap, and then, according to the plan, the successive taps are made feedforwardly.
2. Type II (Do and Adjust). The participants of this type seems just start the procedure and adjust in the course of tapping, probably using (cognitive?) feedback.
3. Type III (Mixture of Type I and II). Type I and II behaviors are not distinct.

In Fig. 3, typical three examples of responses from three participants are shown.

3.2 Error Distribution

Let us look into the error in the number of taps. First we show , in Fig. 4, the distribution of the participants (16 persons in total) with respect to error-making rate. Most (69%) of the participants make less than 10% errors. The relationship between the error and the prescribed number of tapping is summarized in Fig. 5. We can see the abrupt rise in the error rate at *n* = 5. The errors are divided into two types: “too much” and “too few”. In Fig. 5, “too much” errors are shown as white bars, and “too few” errors are as black bars.

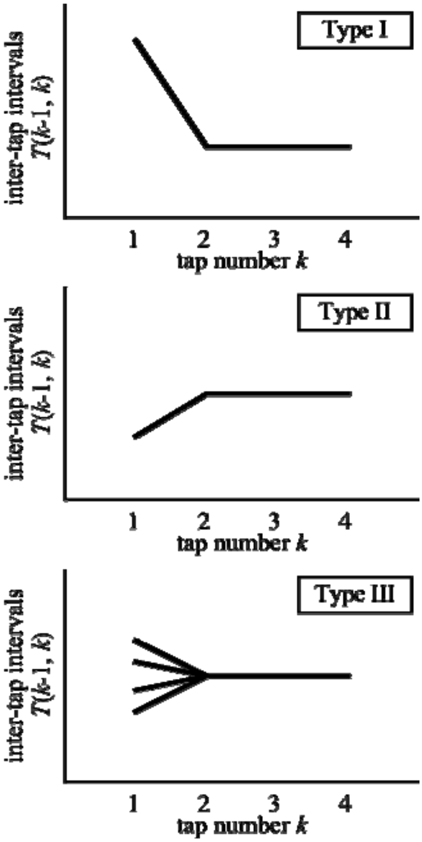
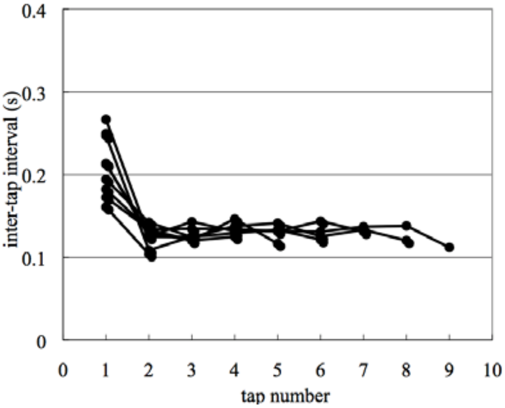
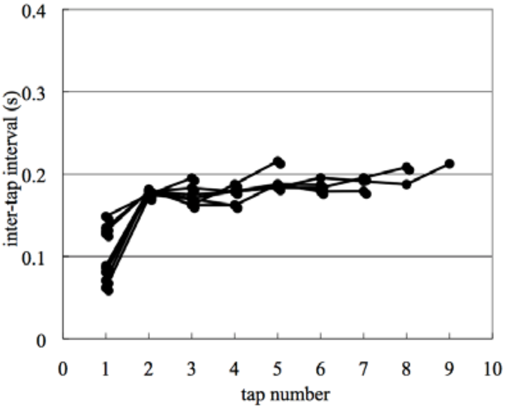


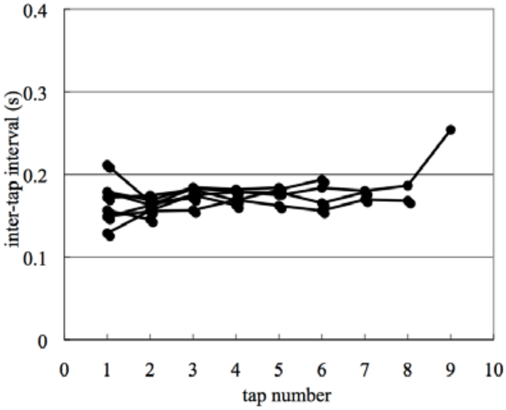
Fig. 2. Three types of time series data. Time series data of the inter-tap intervals are categorized into three types by focusing on the relationship between T(0, 1) and T(1, 2)



(a) An example of Type I.



(b) An example of Type II.



(c) An example of Type III.

Fig. 3. Inter-tap interval vs. tap number for Types I, II and III. Time interval between the tap k–1 and k, T(k–1, k), is plotted as a function of tap number k.

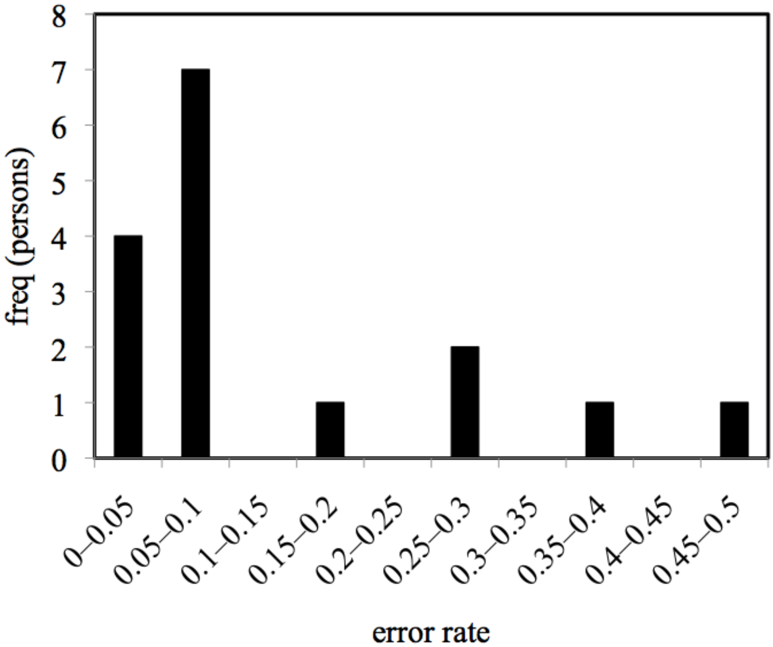


Fig. 4. Distribution of the participants with respect to error rate

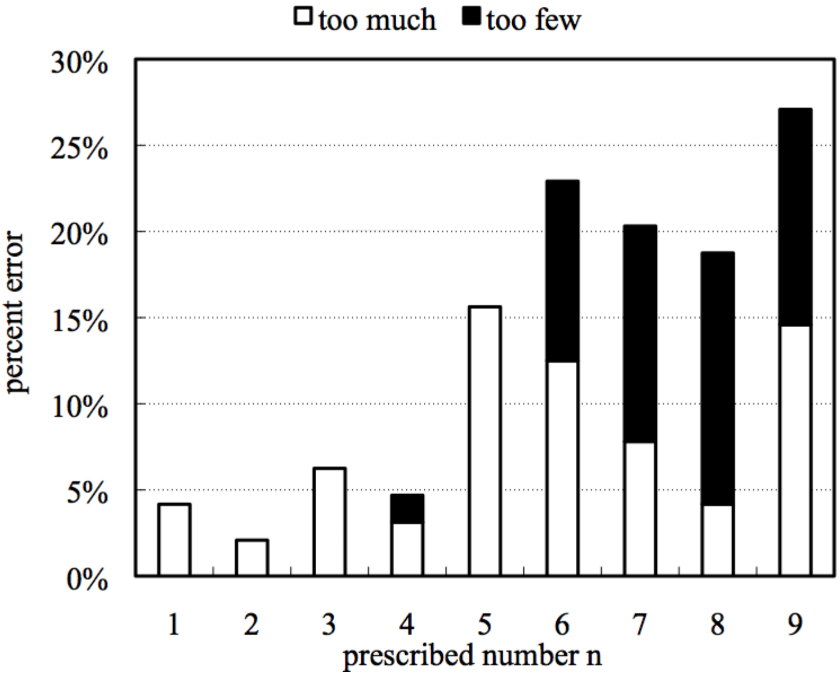


Fig. 5. Error rates averaged over all participants and all trials

4 Summary and Discussion

We investigated time series of inter-tap intervals for the tapping task with time pressure (the participants are to tap prescribed number of times as fast as possible). Looking into the time series of sixteen participants, we tentatively categorized the data into three types. That is, Type I (Plan and Do), Type II (Do and Adjust), and Type III (Mixture of Type I and II). We are just starting the analysis of the data to discuss the validity and the meanings of these types. We are also doing in-depth analysis of the time series of inter-tap intervals in comparison with the simple reaction time and the "Model Human Processor" by Card et al. [4]. The error distribution of Fig. 5 is reminiscent of subitizing-counting modes in enumerating the number of visual objects. That is, in the subitizing range (1–4 in average) you can answer the number of the objects instantly, and in the counting range (5 and over) you have to count up sequentially, hence the time needed to answer the number tends to much longer than in the subitizing range. In enumerating, if there is a time pressure, errors are inevitable in the answer. Thus, we say the present result for error rate is analogous to subitizing-counting characteristics of human behavior. Of course, there is an essential difference between enumerating the objects and multi-tapping of a key. That is, whereas the former is the process of recognition, the latter is of planning and action. We have to analyze the data further in detail with a view to separating the underlying processes. One of other things to consider is doing the same line of experiments by using visual cues instead of auditory ones. We have just gathered the data for the same experiment which uses two different keyboards (difference is in the force needed to press the key), to investigate the effects of the keyboard on the time series of inter-tap intervals and the distribution of errors.

References

1. MacKenzie, I.S.: KSPC (keystrokes per character) as a characteristic of text entry techniques. In: Proc. 4th International Symposium on Human-Computer Interaction with Mobile Devices, pp. 195–210. Springer, Heidelberg (2002)
2. Alston, L., Humphreys, G.W.: Subitization and attentional engagement by transient stimuli. Spatial Vision 17, 17–50 (2004)
3. Camos, V., Tillmann, B.: Discontinuity in the enumeration of sequentially presented auditory and visual stimuli. Cognition 107, 1135–1143 (2008)
4. Card, S.K., Moran, T.P., Newell, A.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Mahwah (1983)

The Number of Trials with Target Affects the Low Prevalence Effect

Fan Yang[1, 2], Xianghong Sun[1], Kan Zhang[1], and Biyun Zhu[1,2]

[1] Institute of Psychology, Chinese Academy of Sciences,
Beijing 100101, China
[2] Graduate University of Chinese Academy of Sciences,
Beijing 100039, China
yangf@psych.ac.cn

Abstract. Wolfe J M. et al found that subject's miss rate increased markedly when target prevalence decreased in simulated X-ray luggage screening task, which was so-called the low prevalence effect. He thought it was caused by shift of observer's decision criteria. But the number of trials with target (NTT) also affected the effect. The present study had two experiments, and there were two blocks in each experiment. Subjects in Exp 1 were in different NTT (20 vs. 100) but the same target prevalence (both 50%); In Exp 2, NTT was the same (both 20) but the target prevalence was different (50% vs. 5%). The results showed that subject's miss rate was mainly changed with NTT, and decision criteria was up to the target prevalence, Wolfe's conclusion was not completely correct.

Keywords: X-ray luggage screening, low prevalence effect, miss rate, visual search.

1 Introduction

The security of transport systems such as airport and subway has attracted more and more people's attention since 9/11 incident, but it is not always reliable, researchers were clear of security check of 15 airports with bombs and guns [1].

X-ray screening especially luggage screening is one of the most important processes in the security of transport systems. It was essentially a visual searching task, "and visual search was a ubiquitous target detection task [2]", but X-ray luggage screening was a little special, its target prevalence was very low, it was about only one time a month per an airport to detect knives or guns in passengers luggage [3]. And there maybe brought a serious problem, observer would miss some targets if the target was too rare. Wolfe J M. et al found that observer's miss error rate increased remarkably (from 7% to 30%) as target prevalence decreased (from 50% to 1%) [4], this was so-called the low prevalence effect, and he considered it was caused by shift of subject's decision criteria when target prevalence decreased. When the target prevalence was very low, the observer always responded with no target, and answering yes would more likely make mistake, he would be very cautious and his

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 126–131, 2011.

decision criteria were shifted to a strongly conservative position [5]; however, the observer could had more false alarms as the target prevalence was very high. And after then, Wolfe had done lots of experiments to cure the effect, but found that it was nearly impossible [5].

We know that the target prevalence is related to NTT and the number of all trials (see Equation 1, N is the number of all trials), and observer's miss rate is equal to the number of trials that targets are missed (Nm) divided NTT (see Equation 2), from Equation 1 and 2, we could get Equation 3. That is to say, observer's miss rate is related to Nm and NTT, or Nm, target prevalence and N. And among these variables, NTT, N and target prevalence can be controlled by experimenter; Nm is up to the observer (e.g., decision criteria) and stimulating materials (e.g., background) and so on. According to Equation 1, only two variables of NTT, N and target prevalence are independent, and the other one is dependent, for example, if N and target prevalence are given, NTT can be calculated. So we should keep NTT or N fixed when researching how target prevalence affects observer's miss rate.

$$\text{target prevalence} = \text{NTT} / \text{N} \tag{1}$$

$$\text{miss rate} = \text{Nm} / \text{NTT} \tag{2}$$

$$\text{miss rate} = \text{Nm} / (\text{ target prevalence} * \text{N}) \tag{3}$$

However, Wolfe just mentioned the target prevalence but ignored NTT and N. in most of Wolfe's experiments about the low prevalence effect, there were 200 trials in high target prevalence (50%) block, NTT was 100; and 1000 trials in low target prevalence (2%) block, NTT was 20. In these two blocks, neither NTT nor N is the same. There would be a problem, if an observer misses both 18 targets in these two blocks, the miss rate in high target prevalence block will be 18%, but the miss rate in low target prevalence would be 90%! This experiment designing would magnify the low prevalence effect.

How NTT and target prevalence affected the miss rate and decision criteria was this study's aim.

2 Methods

2.1 Materials

X-ray pictures used in this study were created in laboratory. Each X-ray picture was in noisy background and contained 18 disturbing objects, which were ordinary things such as watches, cell phones, shoes, glasses, keys, tools (scissors, pincers, etc), toys (cars, tractors, tanks, etc), bottles, cameras and so on. And pictures with target of course had a target, which was a knife or gun. All disturbing objects and targets were black and white, partly transparent and in random rotation, and then overlapped randomly (see figure 1).

Exp 1 and Exp 2 used the same pictures.

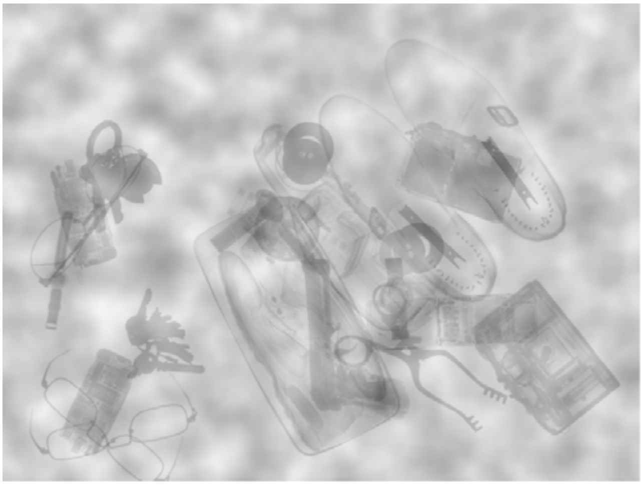


Fig. 1. A picture with a gun

2.2 Participants

15 college students (7 male, 8 female, ages were 18-26 years) were tested in this study, they all reported no history of eye or muscle disorders, and their visual acuity were normal or corrected to normal. 9 students (5 male, 4 female) participated in Exp 1, and 6 students (3 male, 3 female) participated in Exp 2.

2.3 Procedure

Exp 1 and 2 had similar procedure. Subjects were familiar with targets (guns and knives) at first, then practiced searching 10 times (the target prevalence was 50%), and at last, they had a formal screening task which concluded 2 blocks. Subjects were asked to have a rest for 2 minutes between the blocks.

And in Exp 1, the 2 blocks had different NTT (20 vs. 100) and N (40 vs. 200) but the same target prevalence (both 50%, it can be easily figured out by formula 1); however, the NTT were the same (both 20) in Exp 2, but N (40 vs. 400) and target prevalence (50% vs. 5%) were both different. And subjects rest twice in Exp 2, they also had 2 minutes rest in the half of the block with 400 trials besides the rest between blocks.

In each trial, the "+" was on the center of screen for 500 ms at first, and then a X-ray picture showed, subjects should press the key "1" when they found the target; otherwise press the key "2", their responses and reaction time (RT) were recorded. Feedback was given for right or wrong responses in practice, but no feedback in formal searching. And subjects were asked to response as soon as possible in condition of making sure their answers were correct.

2.4 Data Analysis

The trial with no response or that RT less than 200 ms was removed. And the subject that missed far too much times and responded extremely fast was also out of analysis. As a result, one subject's data were eliminated in each experiment, and 8 subjects (4 male, 4 female) were left in Exp 1, 5 subjects (2 male, 3 female) were left in Exp 2.

3 Result

Subjects' miss rates, decision criteria and RT were dependent variables and calculated after experiments. The three variables between blocks in each experiment were compared respectively by Paired-Samples T Test in SPSS 13.

Decision criteria is a psychology variable related to the subject, it increases means that subject becomes conservative, and more likely to consider that there is no target in the picture. The calculating formula in EXCEL 2003 is as follows, decision criteria =- (norminv(hit%)+norminv(false alarm%))/2 [5].

In Exp 1, the two blocks had different NTT (20 vs. 100), same target prevalence (both 50%), miss rate increased markedly ($t = 4.25$, $p = 0.004$) as NTT increased (see figure 2), it indicated that NTT also affected the low prevalence effect even the target prevalence didn't change. Subjects' decision criteria (see figure 3) had no difference ($t = -0.43$, $p = 0.681$), and RT (see figure 4) was also nearly the same ($t = 0.23$, $p = 0.826$).

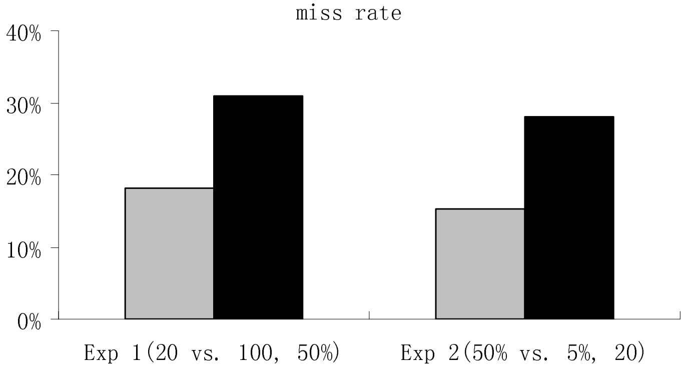


Fig. 2. Subject's miss rate in different NTT or target prevalence

And in Exp 2, target prevalence was different (50% vs. 5%) and NTT was the same (both 20) between the two blocks, the difference of miss rate was not very remarkable ($t = -2.6$, $p = 0.06$), which indicated that target prevalence didn't influence on miss rate very much, the low prevalence effect was not very obvious. And subject's decision criteria ($t = 24.19$, $p < 0.001$) and RT ($t = -3.229$, $p = 0.032$) were remarkable different, the result of Exp 2 was similar with Wolfe's experiments.

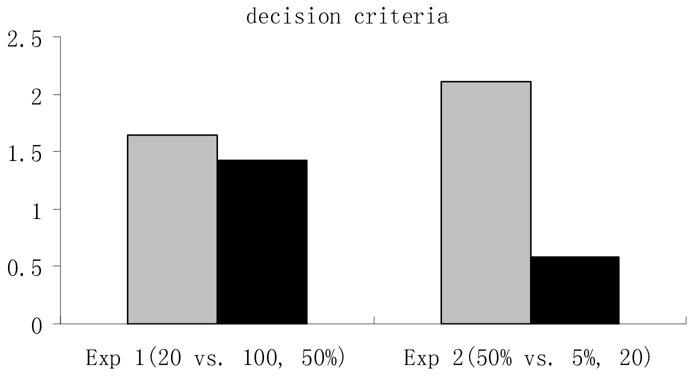


Fig. 3. Subject's decision criteria in different experiment

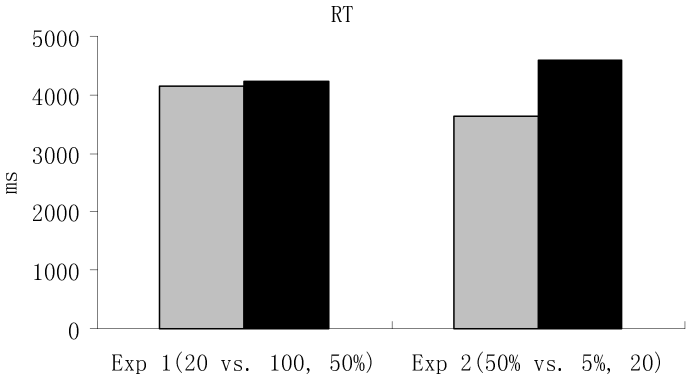


Fig. 4. Subject's RT in different experiment

The detailed data about mean value and standard deviation (SD) of miss rate, decision criteria and RT was in table 1.

Table 1. Mean and SD of miss rate, decision criteria and RT

	Exp 1 NTT = 20 Mean (SD)	Exp 1 NTT = 100 Mean (SD)	Exp 2 50% Mean (SD)	Exp 2 5% Mean (SD)
Miss rate	18.22% (0.08)	30.83% (0.08)	15.24% (0.09)	28% (0.18)
Decision criteria	1.65 (0.98)	1.42 (0.99)	2.16 (0.23)	0.79 (0.56)
RT (ms)	4151 (906)	4215 (474)	3558 (527)	4297 (907)

The results indicated that NTT had more influence on miss rate than target prevalence, but less influence on decision criteria and RT than target prevalence.

4 Discussion

According to formula 2 and 3, miss rate was affected by NTT or the product of target prevalence and N, maybe this was partly the reason that NTT had more influence on miss rate than target prevalence.

The difference of miss rate between blocks in Exp 2 was marginal notable (p=0.06), maybe it will be remarkable if adding more subjects. However, comparing Exp 2 with Wolfe's experiments, equal NTT between different target prevalence blocks would weaken the low prevalence effect. Making NTT the same in different target prevalence blocks, the change of miss rate will be more objective and correct.

References

1. Hall, M.: Airport Screener Missed Weapons. Usa Today (September 23, 2004)
2. Palmer, J., Verghese, P., Pavel, M.: Vision Res., vol. 40, pp. 1227–1268 (2000)
3. Hancock, P.A., Hart, S.G.: Defeating Terrorism: What Can Human Factors/Ergonomics Offer? Ergonomics in Design 10(1), 6–16 (2002)
4. Wolfe, J.M., Horowitz, T.S., Kenner, N.M.: Rare Items Often Missed in Visual Searches. Nature 435(7401), 439–450 (2005)
5. Wolfe, J.M., Horowitz, T.S., Van Wert, M.J., et al.: Low Target Prevalence Is a Stubborn Source of Errors in Visual Search Tasks. Journal of Experimental Psychology: General 136(4), 623–638 (2007)

Part II
Cognitive Aspects of Driving

Facial Expression Measurement for Detecting Driver Drowsiness

Satori Hachisuka[1], Kenji Ishida[1], Takeshi Enya[1], and Masayoshi Kamijo[2]

[1] DENSO CORPORATION, Research Laboratories
500-1, Minamiyama, Komenoki-cho, Nisshin-shi, Aichi-ken, 470-0111, Japan
[2] Interdisciplinary Graduate School of Science and Technology, Shinshu University
3-15-1, Tokida, Ueda-shi, Nagano-ken, 386-8567, Japan
arimitsu@rlab.denso.co.jp

Abstract. This paper presents the method of detecting driver's drowsiness level from facial expressions. Our method is executed according to the following flow: taking a driver's facial image, tracing the facial features by image processing, and classifying the driver's drowsiness level by pattern classification. We found that facial expression had the highest linear correlation with brain waves as the general index of drowsiness during monotonous driving. After analyzing the facial muscle activities, we determined 17 feature points on face for detecting driver drowsiness. A camera set on a dashboard recorded the driver's facial image. We applied Active Appearance Model (AAM) for measuring the 3-dimensional coordinates of the feature points on the facial image. In order to classify drowsiness into 6 levels, we applied k-Nearest-Neighbor method. As a result, the average Root Mean Square Errors (RMSE) among 13 participants was less than 1.0 level. Our method also detected the driver's smile.

Keywords: Facial expression, Facial muscle, Driver drowsiness, Drowsiness detection.

1 Introduction

Although active safety systems in vehicles have contributed to the decrease in the number of deaths occurring in traffic accidents, the number of traffic accidents is still increasing. Driver drowsiness is one reason for such accidents and is becoming an issue. The National Highway Traffic Safety Administration (NHTSA) estimates that approximately 100,000 crashes each year are caused primarily by driver drowsiness or fatigue in the United States [1]. In Japan, attention lapse, including that due to driving while drowsy, was the primary reason for traffic accidents in 2008. The Ministry of Economy, Trade and Industry in Japan reports that the number of such accidents has increased 1.5 times in the 12-year period from 1997 to 2008 [2].

One solution to this serious problem is the development of an intelligent vehicle that can predict driver drowsiness and prevent drowsy driving. The percentage of eyelid closure over the pupil over time (PERCLOS) is one of the major methods for the detection of the driver's drowsiness [3]. We developed a method for the detection

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 135–144, 2011.

of driver drowsiness using the whole facial expression, including information related to the eyes. This method is based on the results of observational analysis. The results of this analysis revealed that features of drowsiness appear on the eyebrows, cheeks, and mouth, in addition to the eyes [4]. The aim of using the facial expression is to detect drowsiness in the early stages, on the basis of the many minute changes in the facial parts. Our goal is to develop an intelligent safety vehicle that can relieve drivers from struggling against drowsiness by detecting their drowsiness and keeping them awake naturally by providing feedback system. In this paper, we discuss a method for detecting driver drowsiness in the early stages. Our method detects drowsiness with accuracy equivalent to that of brain waves, which is the general index of drowsiness. The method categorizes drowsiness into 6 levels using features of facial expression based on the mechanism of facial muscle activities without any attached sensors. We developed the drowsiness detection method with a system comprising a camera set on a dashboard, an image processing algorithm, and a drowsiness detection algorithm.

This paper presents a novel drowsiness detection method and assesses its effectiveness.

2 Early-Stage Drowsiness Detection

The changes in brain waves, especially alpha waves, are one of the indices used to detect changes in the level of drowsiness [5]. Although change in brain waves is an effective index for detecting drowsiness, it is not feasible to apply this index in a vehicle because of the electrodes that are used as contact-type sensors. However, it is recognized in the field of cerebral neuroscience that the facial nerve nucleus is contained in the brain stem, which is defined as an organ of drowsiness [6]. Therefore, we adopted facial expression as the index of drowsiness as an alternative to brain waves. In addition, it is apparent from our experience that we can recognize drowsiness in others from their facial expressions.

In Japan, Kitajima's trained observer rating is a commonly used method for the detection of driver drowsiness on the basis of appearance [7]. The method divides drowsiness into 5 levels with criteria such as "slow blink", "frequent yawning", and so on. Since these criteria are qualitative, the method is not appropriate for automatic detection of drowsiness. Therefore, the quantitative method to detect facial expression was required. To determine the best index as an alternative to brain waves, we examined the correlation between brain waves and other indices such as PERCLOS, heart rate, lane deviation, and facial expression [8]. Those indices do not require the attachment of sensors [3, 9-15]. According to the result, facial expression has the highest correlation with brain waves (correlation coefficient = 0.90) and it detects drowsiness at an earlier stage than other indices. This indicates that facial expression is the most appropriate index to use for the detection of driver drowsiness in the early stages. Therefore, to be able to predict and prevent drowsy driving, the development of a method that detects driver drowsiness from facial expression is necessary.

3 Automatic Drowsiness-Detection Using Facial Expression

It was necessary to solve 3 problems for the development of an automatic drowsiness-detection system:

1. How to define the features of drowsy expression.
2. How to capture the features from the driver's video-recorded facial image.
3. How to estimate the driver's drowsiness index from the features.

Our approaches to solving these problems are explained in this chapter.

3.1 Features of Drowsy Expression

We clarified the particular features of drowsy expression by comparing the facial muscle activities of the waking expression with those of the drowsy expression [16]. We measured 9 facial muscles of each of 17 volunteer participants during the task of monotonous driving in the driving simulator for 1 hour. The left side of Fig. 1 shows the site of 9 facial muscles; inner frontalis, upper orbicularis oculi, lower orbicularis oculi, zygomaticus major, masseter, risorius, upper orbicularis oris, lower orbicularis oris, and mentalis. We divided the reference states of drowsiness into 6 levels, i.e., "Not Sleepy", "Slightly Sleepy", "Sleepy", "Rather Sleepy", "Very Sleepy", and "Sleeping" by adding the "Sleeping" level to Kitajima's trained observer rating scale [7].

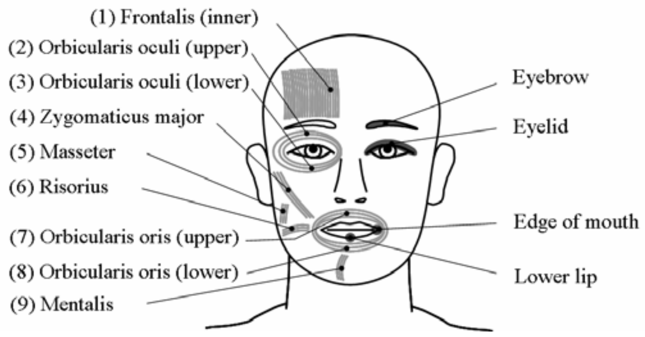
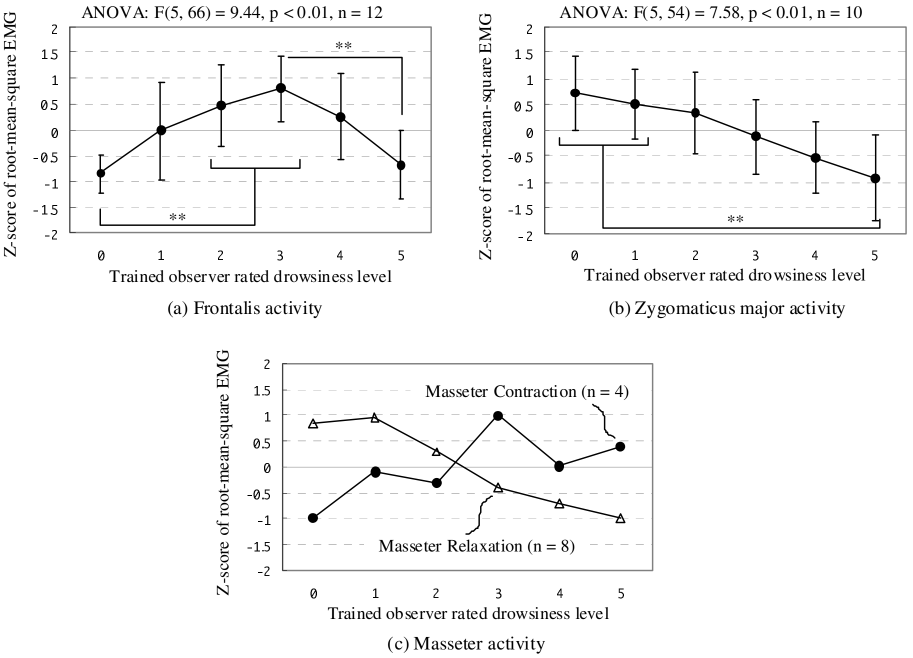

(1) Frontalis (inner)
(2) Orbicularis oculi (upper)
(3) Orbicularis oculi (lower)
(4) Zygomaticus major
(5) Masseter
(6) Risorius
(7) Orbicularis oris (upper)
(8) Orbicularis oris (lower)
(9) Mentalis
Eyebrow
Eyelid
Edge of mouth
Lower lip

Fig. 1. Sites of 9 facial muscles and 4 facial features. Facial muscles were measured by facial electromyograph.

Figure 2 shows the comparison results of the drowsiness levels and the facial muscle activities. The contractions of the frontalis and the relaxation of the zygomaticus major were detected in more than 75 percent of participants, and the relaxation or contraction of the masseter muscle was detected in 82 percent of participants. In addition, contraction of the frontalis, which was detected in 94 percent of participants, was the characteristic expression of resisting drowsiness. This characteristic expression does not appear during the natural drowsy state without any struggle against drowsiness. According to the result, we chose the eyebrows, edges of the mouth, and the lower lip as the facial features related to the frontalis, zygomaticus major, and masseter, respectively, in addition to the eyelids, which are the general features of the drowsiness expression (The right side of Fig.1).



(a) Frontalis activity

(b) Zygomaticus major activity

(c) Masseter activity

Fig. 2. Typical comparison results of the drowsiness levels and the facial muscle activities. The contractions of frontalis and the relaxation of zygomaticus major were detected in more than 75 percent of participants, and the relaxation or contraction of the masseter muscle was detected in 82 percent of participants.

3.2 Image Processing for Measuring Features of Drowsy Expression

We developed a method of image processing for measuring the features of drowsy expression, without any sensor contact, from a driver's video-captured facial image [17]. The method, which is based on the Active Appearance Model (AAM) [18], detects 3-dimensional coordinates of measurement points on the driver's face per frame. Our AAM consists of the specific 2-dimensional model and the generic 3-dimensional model. The specific 2-dimensional model has information relating to the shape and texture of the individual driver's facial image. The generic 3-dimensional model has the 3-dimensional vectors of each measurement points. We developed a method that extracts change in facial expression without individual differences in the shape of each driver's face by using the generic 3-dimensional model. This method is an effective way of detecting the coordinates of the points on the face in the vehicle, which is expected to be driven by an unspecified number of drivers. The process of this method is shown in Fig. 3. First, the specific 2-dimensional model is generated by a captured static facial image of the driver. This process is performed once for each driver. Next, the specific 2-dimensional model is fitted on each frame of the driver's facial image and the 2-dimensional coordinates of the measurement points are output. Finally, the generic 3-dimensional model is aligned based on the 2-dimensional coordinates and the 3-dimensional coordinates of each measurement points are output per frame. We employed the method of steepest descent to the fitting of the specific 2-dimensional model and the aligning of the generic 3-dimensional model.

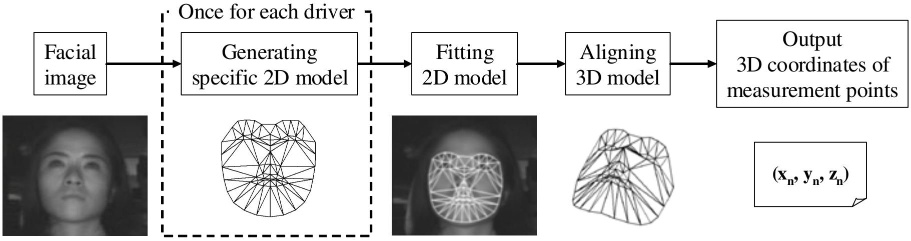


Fig. 3. Flow of the image processing with AAM

3.3 Method of Detecting Drowsiness Level

We adopted 17 points as the measurement objects to detect the drowsy expression. Figure 4 shows the 17 points: 10 points on the right and left eyebrows (5 points on each side), 4 points on the right and left eyelids (2 points on each side), 2 points on the right and left edges of the mouth (1 point on each side), and 1 point on the lower lip. As the features of drowsy expression we used scalar quantities of the change in the 17-point 3-dimensional positions, which were measured from the positions on the waking-state expression. The individual differences of the waking-state expressions are reduced by defining the positions on the waking-state expressions as reference positions. We employed the k-Nearest-Neighbor method, which is one of the pattern classification methods, for detecting the drowsiness level. This decision was based on the result of a preliminary experiment in which we compared the results of the drowsiness levels detected by the trained observer with other estimation methods: multiple regression analysis method, subspace method, and k-Nearest-Neighbor method. The drowsiness level estimated by the k-Nearest-Neighbor method had the highest correlation with the referential drowsiness level as estimated by the trained observer. Our method uses the prebuilt database that consists of the 6-level drowsy expression features of several individuals. The driver's features are compared with the whole database, and the similarities of each comparison are applied to detect the drowsiness level. The similarity-based method is able to detect drowsiness with higher time resolution than the method using trends in the change in the facial expression at a specific time interval, such as 30 seconds [19].

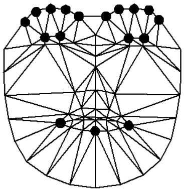


Fig. 4. Seventeen measurement points for detecting drowsy expressions

The every 5-second (150-frame) average features are used as the feature data in this method. The 5-second block time is applied as the bare minimum sampling time for the trained observer rating for facial expressions [7]. According to the averaging, it is possible to detect the difference between "eye closure based on blinking" and "eye closure based on drowsiness", which is difficult to distinguish from a still frame.

We investigated the accuracy of our drowsiness detection. We used the driving simulator in a sound-proof room. Motion system was excluded from the driving simulator to induce drowsiness in the participants efficiently and to measure basic data of participants' drowsy expressions accurately. The driving task was also designed monotonously for the purpose of inducing drowsiness in the participants efficiently. The longitudinal flows of two sine curves, from the top to the bottom of the screen, were projected on the screen. The circle indicating the position of the vehicle from an overhead view was also projected on the screen between the sine curves. We instructed the participants to operate the driving simulator with the steering wheel to maintain their position between the sine curves. The participants' facial images were recorded by the digital video camera (480 x 640 pixels, 30 fps, progressive scan) on the dashboard. The participants were instructed to remain awake during the driving and maintain the same position they would adopt while driving a real vehicle, even if they became drowsy. As a reference for the 6 drowsiness levels, we used the results of ratings for the drivers' recorded facial images from 2 trained observers. The 13 volunteer participants had drivers' licenses and were aged in their 20s to 40s. They were informed of simulator sickness before the experiments and required to sign an informed consent document. During the experiments, at least one examiner observed the participant's appearance from outside the sound-proof room. After the experiment, the participant rested with the examiners for approximately 10-15 minutes.

Drowsiness detection was performed off-line using the leave-one-out cross validation procedure with the features, which were calculated by referring to the 13 participants' recorded facial images. All of the 13 participants fell asleep during the experiment. In the leave-one-out cross validation, data for 12 arbitrarily chosen participants were used as training data and used to detect the drowsiness of the remaining participant, who was excluded from the training data; this was performed repeatedly. The training data consist of the 5-second average features (3-dimensional displacement of 17 points) and the reference of the drowsiness levels, which are labeled on the features. The flow of drowsiness detection is given as follows. The entered driver's features are compared with all of the training data. The top 80 training data, which have a strong similarity to the driver's features, are picked up. The driver's drowsiness level is estimated based on a majority decision of what the referential drowsiness levels are, which are labeled on the 80 training data. We employed the Euclidean distance as the index of similarity between the driver's features and the features of the training data. A small distance indicates strong similarity.

We investigated the effectiveness of the method by comparing the detected drowsiness level with the referential drowsiness level. To detect the facial change based on the drowsiness accurately, we clipped the parts of the video of the participants' facial images before the comparison, and detected the drowsiness levels from those partial videos. The 3 criteria for clipping the partial videos were as follows.

- The facial image of the driver in a front-facing position.
- The facial image without any occlusion such as a steering wheel and/or a hand.
- The facial image without any actions that cause facial change, such as yawning and smiling.

The Root Mean Square Errors (RMSE) of 13 participants are shown in Table 1. These results demonstrate that our method detects the drowsiness with a RMSE of less than 1.0 for 9 participants. The average RMSE among 12 participants is 0.91. On the other hand, RMSE was increased when we used fewer or more feature points than 17 such as 10 points on eyebrows, 4 points on eyelids, or all measurement points on whole face to detect the drowsiness. Therefore, 17 points, which described in chapter 3.3 (FIg. 4), were the best features for our drowsiness-detection method. In addition, we found that if the participant didn't fall asleep during the examination, and its referential drowsiness levels were under level 2 ("Sleepy"), our method indicated the drowsiness levels under level 3 ("Rather Sleepy"). In this paper, this case occurred in only one participant aside from 13 participants.

Table 1. Root Mean Square Errors of drowsiness detection

Participant #	Root Mean Square Error (RMSE)
1	1.06
2	0.90
3	1.14
4	0.91
5	0.78
6	1.11
7	0.71
8	1.00
9	0.82
10	0.81
11	0.93
12	0.77
13	0.85
Average	0.91
SD	0.14

4 Cancellation of Awake Expressions

As we described in the previous chapters, we demonstrated that our method could detect the drowsiness level, when the drowsy expression was input. On the other hand, it is well-known that the facial expressions are divided into 6 global common categories [20] or 9 psychological categories [21]. Since our method only focused on the drowsiness category and the classification of the level, we additionally investigated the result when the facial expression, which was not related to the drowsy expression, was input to our method. We applied "smile" with the entirely awake state as the input expression. The smiling facial images of 4 volunteer participants were recorded by the same camera and the same setup as mentioned in chapter 3.3. The participants consisted of 2 males and 2 females, and were aged in their 20s to 40s. All participants were entirely awake. We input 10-second (300-frame) facial images per participant to our method. As shown in the "Normal output" column on Table 2, the output results were drowsiness level 3 ("Sleepy") or 4 ("Very Sleepy"), respectively instead of the level 0 ("Not Sleepy"). As a feasible method in a real vehicle, these output results must be level 0 or "smile", when a driver smiles in the entirely awake state. The reason of these false detections may be that the facial features such as narrowed eyes and opened mouth were similar to those of drowsy expression.

For the purpose of making a categorical distinction between drowsiness and smile, we examined the difference among the 3-dimensional displacement of 17 facial points. We found that y-axis displacement of the edges of mouth were the most and the second quantity among the facial points. In addition, the threshold of the quantity was over 1.5 (non unit of quantity required) in the 3-dimensional method in this paper. We attached a filter to our method for distinguishing "smile" from "drowsiness" when the displacements of 3-dimensional feature points and the threshold meets the conditions mentioned above. Due to the effect of this filter, we could detect "smile" with 100 percent accuracy among 4 volunteer participants' facial expressions (Table 2). In the similar way, we confirmed the possibility of detecting "speech" with the variance of y-axis displacement of the feature point on lower lip in a few participants.

Table 2. Detected Drowsinesslevel and Expression

Participant #	Normal output		Filtered output	
	0 - 5 [sec]	5 - 10 [sec]	0 - 5 [sec]	5 - 10 [sec]
1	Level 3	Level 4	Smile	Smile
2	Level 3	Level 3	Smile	Smile
3	Level 3	Level 4	Smile	Smile
4	Level 3	Level 4	Smile	Smile

5 Conclusion

In this paper, we presented the driver's drowsiness detection method using facial expression, and we established the effectiveness of this method experimentally. Our method is executed according to the following flow: taking a driver's facial image, tracing 17 feature points by image processing, and rating the driver's drowsiness according to a 6-level scale from the features by pattern classification. The results of the drowsiness detection correspond to the drowsiness reference as estimated by a trained observer with an average RMSE of less than 1.0 level among 13 participants. The distinguishing feature of our method is that it uses 17 facial features based on the activities of facial muscles. In addition, we attached the filter to our method for distinguishing "smile" from "drowsiness" when the facial expression, which was not related to the drowsy expression, was input. This filter effectively detected "smile" with 100 percent accuracy among 4 participants.

The limitations of this paper were the reality of driving environment and the number of the participants. In future work, we will verify practical effectiveness of our drowsiness detection method using motion-based driving simulator and/or real car. In that phase, we will have to distinguish drowsy expression from complexly-mixed expressions. Because, there is a possibility to appear the mixed expressions on drivers' faces, such as "smiling with drowsiness", "speaking with drowsiness" and so on, during the natural expressions in real car. Additionally, we will increase the number of participants in our experiments and develop the effective training data for detecting drowsiness of a large number of drivers.

It is also necessary to develop a feedback system to achieve an intelligent safety vehicle that can relieve drivers struggling against drowsiness. We have now started to develop a feedback system that keeps the driver awake effectively and naturally. In addition, the integration of personal verification into our method will lead to the development of a highly precise drowsiness-detection method.

References

1. U.S. Department of Transportation: Intelligent Vehicle Initiative 2002 annual report. `http://www.itsdocs.fhwa.dot.gov//JPODOCS/REPTS_TE//13821.pdf`
2. The Ministry of Economy, Trade and Industry: Technological Strategy Map (2009), `http://www.meti.go.jp/policy/economy/gijutsu_kakushin/kenkyu_kaihatu/str2009/7_1.pdf` (in Japanese)
3. Wierwille, W.W., Ellsworth, L.A., Wreggit, S.S., Fairbanks, R.J., Kirn, C.L.: Research on Vehicle-Based Driver Status/Performance Monitoring; Development, Validation, and Refinement of Algorithms For Detection of Driver Drowsiness: National Highway Traffic Safety Administration Final Report, DOT HS 808 247 (1994)
4. Ishida, K., Ito, A., Kimura, T.: A Study of Feature Factors on Sleepy Expressions based on Observational Analysis of Facial Images. Transactions of Society of Automotive Engineers of Japan 39(3), 251–256 (2008) (in Japanese)
5. Eoh, H.J., Chou, M.K., Kim, S.H.: Electroencephalographic study of drowsiness in simulated driving with sleep deprivation. International Journal of Industrial Ergonomics 35, 307–320 (2005)
6. Saper, C.B., Chou, T.C., Scammell, T.E.: The sleep switch: hypothalamic control of sleep and wakefulness. TRENDS in Neurosciences 24(12), 726–731 (2001)
7. Kitajima, H., Numata, N., Yamamoto, K., Goi, Y.: Prediction of Automobile Driver Sleepiness (1st Report, Rating of Sleepiness Based on Facial Expression and Examination of Effective Predictor Indexes of Sleepiness). The Japan Society of Mechanical Engineers Journal (Series C) 63(613), 93–100 (1997) (in Japanese)
8. Ishida, K., Hachisuka, S., Kimura, T., Kamijo, M.: Comparing Trends of Sleepiness Expressions Appearance with Performance and Physiological Change Caused by Arousal Level Declining. Transactions of Society of Automotive Engineers of Japan 40(3), 885–890 (2009) (in Japanese)

9. Sato, S., Taoda, K., Kawamura, M., Wakaba, K., Fukuchi, Y., Nishiyama, K.: Heart rate variability during long truck driving work. Journal of Human Ergology 30, 235–240 (2001)
10. Kozak, K., Pohl, J., Birk, W., Greenberg, J., Artz, B., Blommer, M., Cathey, L., Curry, R.: Evaluation of lane departure warnings for drowsy drivers. In: Proceedings of the Human Factors and Ergonomics Society 50th annual meeting, pp. 2400–2404 (2006)
11. Liu, C.C., Hosking, S.G., Lenne, M.G.: Predicting driver drowsiness using vehicle measures: Recent insights and future challenges. Journal of Safety Research 40, 239–245 (2009)
12. Ueno, A., Manabe, S., Uchikawa, Y.: Acoustic Feedback System with Digital Signal Processor to Alert the Subject and Quantitative Visualization of Arousal Reaction Induced by the Sound Using Dynamic Characteristics of Saccadic Eye Movement: A Preliminary Study. In: IEEE Engineering in Medicine and Biology 27th Annual Conference, China, pp. 6149–6152 (2005)
13. Bergasa, L.M., Nuevo, J., Sotelo, M.A., Barea, R., Lopez, M.E.: Real-Time System for Monitoring Driver Vigilance. IEEE Transactions on Intelligent Transportation Systems 7(1), 63–77 (2006)
14. Ji, Q., Zhu, Z., Lan, P.: Real-Time Nonintrusive Monitoring and Prediction of Driver Fatigue. IEEE Transactions on Vehicular Technology 53(4), 1052–1068 (2004)
15. Arimitsu, S., Sasaki, K., Hosaka, H., Itoh, M., Ishida, K., Ito, A.: Seat Belt Vibration as a Stimulating Device for awakening Drivers. IEEE/ASME Transactions on Mechatronics 12(5), 511–518 (2007)
16. Ishida, K., Ichimura, A., Kamijo, M.: A Study of Facial Muscular Activities in Drowsy Expression. International Journal of Kansei Engineering 9(2), 57–66 (2010)
17. Kimura, T., Ishida, K., Ozaki, N.: Feasibility Study of Sleepiness Detection Using Expression Features. Review of Automotive Engineering 29(4), 567–574 (2008)
18. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 681–685 (2001)
19. Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., Movellan, J.: Automated Drowsiness Detection For Improved Driving Safety. In: 4th International conference on Automotive Technologies, Turkey, pp. 13–14 (2008)
20. Ekman, P., Friesen, W.V.: Unmasking the Face. Prentice Hall, New Jersey (1975)
21. Russell, J.A., Weiss, A., Mendelsohn, G.A.: Affect Grid: A Single-Item Scale of Pleasure and Arousal. Journal of Personality and Social Psychology 57(3), 493–502 (1989)

Estimation of Driver's Fatigue Based on Steering Wheel Angle

Qichang He, Wei Li, and Xiumin Fan

Shanghai Key Lab of Advanced Manufacturing Environment, School of Mechanical Engineering, Shanghai Jiaotong University, Shanghai, China, 200030

Abstract. Driver's fatigue has been verified as a major factor in many traffic accidents. The estimation of driver's vigilance by steering wheel angle is good way because it is a non-invasive method compared with EEG. An adaptive vigilance estimation methodology based on steering wheel angle information is proposed. The sample data classification index is built from EEG and PVT information of ten driver's virtual driving experiment on driving simulator. According to the geometry information of road centerline and the location of the automobile center, a new algorithm is proposed to compute the lane deviation. The correlation coefficient between steering wheel angle and lane deviation are computed, and the results show that their correlation level is 0.05. Based on the steering wheel angle, the driver fatigue evaluation model is established by the Bayesian Network (BN). The structure and parameters for BN model are determined after adaptive training. The experiment results verified that this model is effective to identify driver's fatigue level.

Keywords: Driver fatigue; Steering wheel angle; Lane deviation; Bayesian Network model.

1 Introduction

Driver fatigue is a major factor attribute to traffic accident, and the statistics show that traffic accidents caused by driver drowsy accounts for about 20% of the total number of accidents, and more than 40% of serious traffic accident [1]. It has great safety significance to detect driver fatigue level quickly and efficiently.

Many research results have been achieved in detecting driver fatigue, such as physiological signal [2-4], driver's face expression [5-9] and so on. However, these methods have obvious shortcomings: the acquisition of physiological signal requires putting sensor on the driver's body which increases the cost and makes the driver uncomfortable especially in long-time driving; The detection of driver facial expression is influenced by the illumination, especially when the driver wear glasses; Also the information of eyelid movement (Perclos[5]) is hard to obtain in the high bright condition.

At present, driver's manipulation signal attracts more attention in the field of driver fatigue detection. Skipper, etc. [10] found the maximum and MSE (Mean Square Error) of lane deviation in drowsy state is larger than alert state. Siegmund et al [11] constructed three weight functions based on steering wheel angle in time domain,

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 145–155, 2011.

frequency domain and amplitude domain respectively, and then established the index function called SED (Subjective Evaluation of Drowsiness) to evaluate the fatigue level. The experiment results show that SED is larger in drowsy state. Based on the analysis of steering wheel angle in time domain, Eskandarian et al [12] found that if the steering wheel angle is changed large firstly, and then changed little in a certain period time which means the driver is mostly in drowsy state.

From the results of above research, the lane deviation and steering wheel angle can be used to evaluate the driver fatigue level. The advantages of these methods are that it is little affected by the illumination and convenient for the driver because of its non-intrusive detection. However, the accuracy of these non-intrusive methods is lower. In order to improve its accuracy, the Bayesian Network (BN) method is used to build the fatigue level evaluation model in this paper.

The structure of this paper is organized as follows: firstly, the data acquisition process is introduced; secondly, a general algorithm is present to compute the lane deviation and eliminate the road curvature; then, the methods to separate the sample data into alert and drowsy state are provided; after that, the model to evaluate the driver fatigue level is established with the help of the BN method.

2 Method

2.1 Apparatus

A self-developed driving simulator (Fig.1) was used to detect driver fatigue. The hardware of driving simulator includes the Logitech steering wheel called MOMO force feedback racing wheel, Logitech camera and so on. The software of driving simulator is composed of the scene rendering system, the audio rendering system, the automobile dynamics model and the video capture system and so on. The Logitech camera is installed on the dashboard to collect driver's facial expression during driving. The driver's EEG signal is collected by NicoletOne Ambulatory EEG (Fig.2).

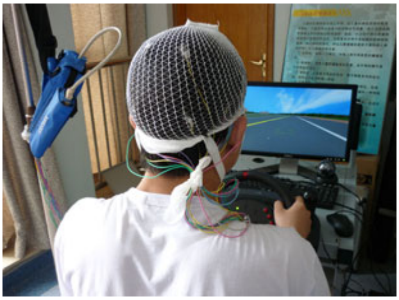


Fig. 1. The driving simulator and virtual driving scene

Fig. 2. NicoletOne Ambulatory EEG

2.2 Subjects

Ten healthy male drivers ranging from 22 to 35 old (Age: 28.1 ± 3.6) were enrolled in this experiment. They all had a legal driver license and normal sleep-wake habits. Also they must have good sleep quality and no physical barrier before the experiment.

2.3 Experiment Arrangement

The experiment is divided into three periods: dawn (00:30-06:00), morning (8:00-11:30), noon (12:30-15:30). Drivers are instructed to drive at 80 km/h for straight line section and 40km/h for curve section. The virtual driving scene is monotonous that make the driver tend to become drowsy. The road is 100 km long composed of 4 lanes with 3.5 meters wide. Drivers are required to complete all the tasks and make proper manipulation to ensure safe driving. Before the experiment, all the drivers have a chance to be familiar with the experiment procedure.

3 Data Analysis

3.1 Lane Deviation

In this virtual driving scene, there are two kinds of road centerlines (curve and straight line), so the lane deviation is computed separately. Fig.3 shows a part of road centerline. AB, CD, EF, GH is the straight line sections, and BC, DE, FG belong to the curve sections.

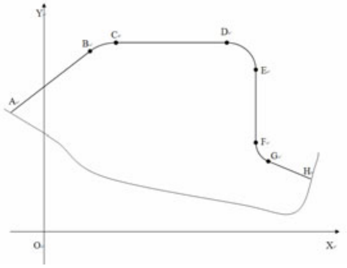
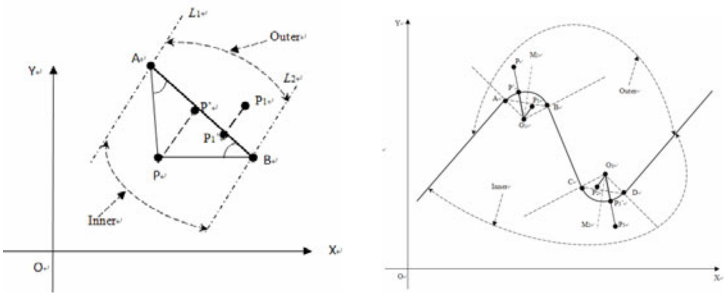


Fig. 3. A part of road centerline in driving scene



(a) straight line

(b) curve section

Fig. 4. The identified algorithm

Fig.4 (a) shows the identified algorithm for straight line. A, B is the two endpoints of one straight line section. L_1, L_2 is the boundary line of AB section. If the point P, P_1 located between line L_1 and line L_2, the point belongs to AB section. Taking the point P for example, the angle of PAB and PBA are no more than 90 degree, so the cosine of PAB and PBA are no less than zeros.

$$\text{As}\begin{cases}\overrightarrow{PA}\bullet\times\overrightarrow{BA}=\left|PA\right|\bullet\left|BA\right|\bullet\cos(PAB)\\ \overrightarrow{PB}\bullet\times\overrightarrow{AB}=\left|PB\right|\bullet\left|AB\right|\bullet\cos(PBA)\end{cases}\tag{1}$$

$$\text{Then } S_{straight}=(\overrightarrow{PA}\bullet\times\overrightarrow{BA})\times(\overrightarrow{PB}\bullet\times\overrightarrow{AB})\geq 0$$

If S_{straight} is no less than zero, the point P belongs to the AB section; otherwise it doesn't belong to AB.

Fig.4 (b) shows the identified algorithm for curve line. O_1 is the center of arc AB. A, B is the endpoints of arc; P is the automobile position. O_1M_1 is the perpendicular bisector of AO_1B. AO_1, BO_1 is the boundary line. If the point located between these two lines, it belongs to AB curve section. Take the point P for example, it can be seen from Fig.4 (b) that the angle of PO_1M_1 is no larger than AO_1M_1, so the cosine of PO_1M_1 is no smaller than that of AO_1M_1.

$$\begin{gathered}\begin{cases}\overrightarrow{M_1O_1}\bullet\times\overrightarrow{PO_1}=\left|M_1O_1\right|\bullet\left|PO_1\right|\bullet\cos(M_1O_1P)\\ \overrightarrow{M_1O_1}\bullet\times\overrightarrow{AO_1}=\left|M_1O_1\right|\bullet\left|AO_1\right|\bullet\cos(M_1O_1A)\end{cases}\Rightarrow\\ S_{curve}=\overrightarrow{M_1O_1}\bullet\times\overrightarrow{PO_1}-\overrightarrow{M_1O_1}\bullet\times\overrightarrow{AO_1}\times\frac{\left|PO_1\right|}{\left|AO_1\right|}\geq 0\end{gathered}\tag{2}$$

If the value of S_{curve} is no less than zero, the point belongs to current AB curve section.

Considering the two kinds of road centerline, the lane deviation can be computed as Eq. (3).

$$\begin{gathered}D_L=\frac{x(y_2-y_1)+y(x_1-x_2)-x_1y_2+x_2y_1}{\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}}\quad\text{straight\_line}\\ D_C=\sqrt{(x_a-x_0)^2+(y_a-y_0)^2}-\sqrt{(x-x_0)^2+(y-y_0)^2}\quad\text{curve}\end{gathered}\tag{3}$$

Where: D_L *and* D_C is the lane deviation in the straight line section and curve section respectively; x_1, y_1, x_2, y_2 is the coordinate value of two endpoints in the straight section; *x*, *y* is the current coordinate location value of automobile center; x_0, y_0 is the coordinate value of curve center; x_a, y_a is the coordinate value of one endpoints of curve.

In the virtual driving scene, all the road centerlines is constructed to form a closed loop, so the point can be located in the inner or outer of the closed loop. For the straight line section in Fig.4 (a), P is the inner point and P_1 is the outer point, so the lane deviation of P is positive and P_1 is negative. For the curve section in Fig.4 (b), P and P_2 are in the outer section, P_1 and P_3 are in the inner section. However the lane deviation for P /P$_3$ is positive and P$_1$/P$_2$ is negative which is different with straight line. In order to ensure the sign of the lane deviation can be acted as the judgment parameter to distinguish the inner section and outer section, C$_{sign}$ is introduced to ensure the sign of lane deviation for P/P$_2$ be negative and P$_1$/P$_3$ be positive (Eq.(4)).

$$C_{sign} = x_{O_1}(y_B - y_A) + y_{O_1}(x_A - x_B) - x_A y_B + x_B y_A \tag{4}$$

So the lane deviation can be computed as Eq. (5)

$$\begin{aligned} D_L &= \frac{x(y_2 - y_1) + y(x_1 - x_2) - x_1 y_2 + x_2 y_1}{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} \quad \text{straight\_line} \\ D_C &= (\sqrt{(x_a - x_0)^2 + (y_a - y_0)^2} - \sqrt{(x - x_0)^2 + (y - y_0)^2}) \times C_{sign} \quad \text{curve} \end{aligned} \tag{5}$$

3.2 The Road Curvature Elimination

In the straight line section, the steering wheel angle only reflects the driver's steering adjustments, which would be affected by the road curvature in the curve section. It must be eliminated in the curve section to ensure the steering wheel angle only reflect the steering adjustments.

The curve section is divided into three parts: the enter part, the middle part and the exit part. The data of steering wheel angles are divided into groups (5 data for one group). The algorithm of road curvature elimination is described as follows:

1. *The enter part*

When the automobile is from straight line section to curve section, the steering wheel angle is from small to large until it waves around a special value. The identification of this procedure is as follow:

$$\begin{cases} std(A_{s,i}) > A_t \\ std(A_{s,i+1}) < A_t \end{cases} \quad i = 1, 2, \cdots, r \tag{6}$$

Where: std is the function to compute the standard error of data; A_t is the provided angle which used to distinguish the steering wheel angle belongs to the transition part (the enter or exit part of curve section) or not.

Then the steering wheel angles in the group between *first* and r^{th} belong to the enter part of curve section. The curvature in this part can be eliminated as Eq. (7)

$$A_{s,i,j} = \frac{(A_{s,i,j} - \frac{1}{5}\sum_{j=1}^{5} A_{s,i,j})}{\frac{1}{5}\sum_{j=1}^{5} A_{s,i,j}} \quad i = 1, 2, \cdots, r \tag{7}$$

2. *The middle part*

While the automobile is located in the middle part of curve section, the steering wheel angle changes at a small range. The identification algorithm of this part is as follow:

$$\begin{cases} std(A_{s,i}) < A_{\mathrm{t}} \\ std(A_{s,i+1}) > A_{\mathrm{t}} \end{cases} \quad i = r+1, r+2, \cdots, s \tag{8}$$

It can be seen from above that the driving steering wheel angle between (r+1)th and sth group belong to the middle part of curve section.

The curve curvature in this part can be eliminated by the method as Eq. (9).

$$A_{s,i,j} = A_{s,i,j} - \frac{1}{5}\sum_{j=1}^{5} A_{s,i,j} \quad i = r+1, r+2, \cdots, s \tag{9}$$

3. The exit part

While the automobile is driving from curve to straight line section, the steering wheel angle is from large to small until it waves around a special value. The curvature elimination in these groups can be eliminated as Eq. (10).

$$A_{s,i,j} = \frac{(A_{s,i,j} - \frac{1}{5}\sum_{j=1}^{5} A_{s,i,j})}{\frac{1}{5}\sum_{j=1}^{5} A_{s,i,j}} \quad i = s+1, s+2, \cdots, n \tag{10}$$

Fig. 5 shows the steering wheel angle before and after the road curvature. It can be seen that the steering wheel angle after curvature elimination only reflects the steering adjustments.

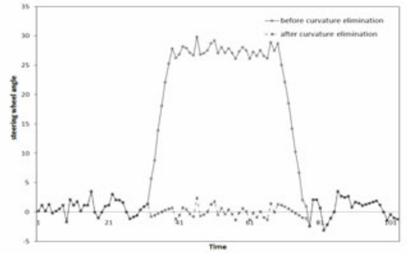


Fig. 5. The road curvature elimination

3.3 The Sample Data Classification

In this study, the sample data is classified into two fatigue level (alert & drowsy) with reference to the EEG, PVT and driving video.

EEG. As the EEG signal is regarded as the gold index to evaluate the driver fatigue level, it is introduced to be an index to classify the sample data. The driver's EEG signal is collected by NicoletOne Ambulatory with the sampling frequency of 200 Hz. The EEG signals are transformed to frequency domain by FFT method. It can be separated into 4 wave bands: δ (0.3-3.5 Hz),

θ(4-8Hz), α(8-13Hz), β(14-20Hz). A significant increase of θ and α activity, and a slight decrease of β activity has been indicate the drowsy state. [2,13]. Based on this results, the power spectrum ratio of θ and α relate to β called $R_{(\theta+\alpha)/\beta}$ is considered as the index to evaluate the driver fatigue level.

$$R_{(\theta+\alpha)/\beta} = \frac{A_\theta + A_\beta}{A_\beta} \tag{11}$$

Where: $A_\theta, A_\alpha, A_\beta$ is the power spectrum of θ, α and β respectively.

The EEG signal is grouped according to a given time interval (240 seconds). Then the mean power spectrum of each band is computed. The index for detect driver fatigue is computed as Eq. (12).

$$\begin{aligned} R_{eeg} &= \min(\frac{1}{A_{eeg}}(\lambda_1 \rho_1 \times mean(\frac{A_{\theta,i} + A_{\alpha,i}}{A_{\beta,i}}) + \\ &\lambda_2 \rho_2 \times std(\frac{A_{\theta,i} + A_{\alpha,i}}{A_{\beta,i}})),1) \quad i = 1, 2, \cdots, n \end{aligned} \tag{12}$$

Where: min is the function to compute the minimum value ; mean is the function to compute the mean value; std is the function to compute the MSE; A_{eeg} is the normalized results of $R_{(\theta+\alpha)/\beta}$; ρ_1, ρ_2 is the factor with reference to the individual differences; λ_1, λ_2 is the weight coefficient.

PVT. PVT test is proved to be an efficient method to evaluate the driver fatigue level. The score of driver's PVT test is composed of two parts: the response time and the accuracy of judgment. The score of PVT test for driver fatigue level is calculated by Eq. (13).

$$R_{PVT} = \min(w \bullet (w_1 \bullet \frac{t_R}{t_{SR}} + w_2 \bullet a_{PVT}),1) \tag{13}$$

Where: R_{PVT} is the score of PVT test for driver fatigue level; t_R is the response time; tSR is the mean response time when the driver in alert state; aPVT is the accuracy of judgment; w is a weight coefficient to indicate individual differences; w1, w2 is the weight coefficient.

Besides the EEG signal and PVT test, the driving video is as an additional index to evaluate the fatigue level. The score of driving video for fatigue level can be computed refer to the Perclos [5]. The final index of driver fatigue level is calculated by Eq. (14).

$$R_f = \rho_p \times R_{pvt} + \rho_e \times R_{eeg} + \rho_v \times R_v \tag{14}$$

Where: R_f is the final scores of driver fatigue level; R_V is the score of video assessment; ρ_p, ρ_e, ρ_v is the weight coefficient. In this paper, ρ_p=0.5, ρe=0.35, ρv=0.15.

3.4 The Correlation of Steering Wheel Angle and Lane Deviation

The lane deviation and steering wheel angle collected in driving experiment are classified into alert and fatigue level. The correlation coefficient between lane deviation and steering wheel angle is computed under these two fatigue levels. These two types of data are processed to eliminate the noise and road curvature before the computation. Tab.1 shows that the steering wheel angle has high correlation with lane deviation. As the steering wheel angel is collected easily, it is chose to be the index to evaluate the driver fatigue level in this paper.

Table 1. The correlation coefficient

Fatigue level	alert	drowsy
Correlation Coefficient	0.3568	0.3198

4 Results

4.1 Data Processing

In order to meet the demand of BN model and reflect the distribution of steering wheel angle in a certain time interval, the steering wheel angle is discrete and normalized.

Discretization. The steering wheel angle is discrete by Eq. (15), Eq. (16).

$$\text{Initialization}: \ Q_{a,i} = 0, \ \ i = 1, \cdots, n \tag{15}$$

$$Q_{a,e} = 1, \quad e = \max(\min(floor(\frac{a - a_m}{L_a}) + \frac{n}{2} + 1, n), 1) \tag{16}$$

Where: Q_a is the vector after the data is discrete; a is the steering wheel angle; a_m is the mean steering wheel angle; floor is a function to choose the integer round towards minus infinity; L_a is the length of every interval. Choosing different value for L_a, Q_a can be calibrated for different driving behavior. Large values of L_a are used for driver with large steering wheel adjustment behavior while they are drowsy.

Normalization. After the data being discrete, all data in one time interval should be summed up, and then the data have to be normalized by Eq. (17).

$$Q_{a,i} = \frac{\sum_{j=1}^{20} Q_{a,ij}}{\max(\sum_{j=1}^{20} Q_{a,ij})}, \ i = 1, \cdots, n \tag{17}$$

Where : $Q_{a,i}$ is the normalized result of steering wheel angle in the i^{th} time interval.

4.2 BN Model

BN model is a probabilistic graphical model that represents a set of random variables and their conditional dependences via a directed acyclic graph (DAG). BN has several advantages for data analysis: handle situation where some data are missing; gain understanding about a problem domain and predict the consequences of intervention; provide an ideal representation for combining prior knowledge and data; offer an efficient and principled way for avoiding the over fitting of data [14].

It can be seen from Fig.6 that the steering wheel angle is almost the normal distribution. Considering steering wheel angle being the normal distribution, a new BN model is proposed to detect driver fatigue level based upon Gaussian mixture models (GMM) which was usually used as classification tool [16].

The model is a two-class, two component mixture model: class 1 for alert state and class 2 for drowsy state. Its structure is described in Fig.7.

In this paper, the output (Node 3, the dimensional feature) and the class (Node 1, alert and drowsy) are observed. The type of Conditional probability Distribution (CPD) of driver fatigue level and the steering wheel angle is tabular and Gaussian respectively.

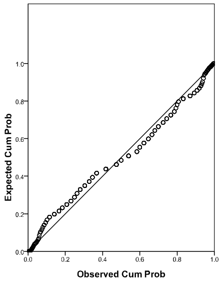


Fig. 6. The P-P plot of steering wheel angle

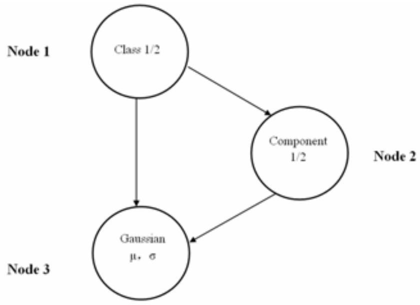


Fig. 7. The graph structure of BN

The jtree_inf_engine[p] is chose as the inference engine. The node size of driver fatigue level and steering wheel angle is 2 and 12 respectively. 500 samples are utilized to train the BN model. The iteration number of EM algorithm is set to 10 and the stopping criterion is 0.01. After training, the parameters of BN are determined. Based on the BN model, the driver fatigue level can be evaluated. Fig.8 shows the training procedure.

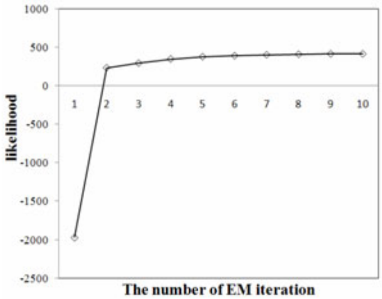


Fig. 8. The training of GMM

Table 2. The evaluation result

Original data/Group	Test Result		
	alert	drowsy	Correct ratio/%
50(alert)	41	9	82
50(drowsy)	12	38	76

5 Discussion and Conclusion

After the BN model is established, another 50 experiment data from alert and drowsy state are chose to evaluate the accuracy. Tab.2 shows the evaluation result. The model identified 41 out of a total of 50 alert samples and 38 out of a total 50 drowsy samples. So the accuracy of the model is about 79%. In this paper, the steering wheel angle is chose as the index to evaluate the driver fatigue level based on BN method. The experiment results show that the model is efficient and practical for fatigue level evaluation. However, the manipulation of steering wheel is greatly influenced by the individual habits, so the model that only considering this source data is not precise enough. The model incorporated multi-sources will be studied in the future.

References

[1] Mao, Z., Chu, X.-m., Yan, X.-p., et al.: Advances of fatigue detecting technology for drivers. China Safety Science Journal 15(3), 108–112 (2005)

[2] Jap, B.: Using EEG spectral components to assess algorithms for detecting fatigue. Expert Systems with Applications 36(2), 2352–2359 (2009)

[3] Lal, S.K., Craig, A.: A critical review of the psychophysiology of driver fatigue. Biological Psychology 55(3), 173–194 (2001)

[4] Eoh, H.J., Chung, M.K., et al.: Electroencephalographic study of drowsiness in simulated driving with sleep deprivation. Int. J. Of Industrial Ergonomics 35(4), 307–320 (2005)

[5] Luis, M.B., Miguel, A.S.: Real-time system for monitoring driver vigilance. IEEE Transactions on Intelligent Transportation Systems 7(1), 63–77 (2006)

[6] Rongben, W., et al.: Monitoring mouth movement for driver fatigue or distraction with one camera, Washington, DC, United states, pp. 314–319 (2004)

[7] Morad, Y., Barkana, Y., Zadok, D., et al.: Ocular parameters as an objective tool for the assessment of trunk drivers fatigue. Accident analysis and prevention 41(4), 856–860 (2009)

[8] Mohanty, M., Mishra, A., Routray, A.: A non-rigid motion estimation algorithm for yawn detection in human drivers. Int. J. of Computational Vision and Robotics, 89–109 (2009)

[9] Gu, H., Ji, Q., Zhu, Z.W.: Active facial tracking for fatigue detection. In: Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision, Orlando, pp. 137–142 (2002)

[10] Skipper, J., Wierwille, W., Hardee, L.: An investigation of low-level stimulus-induced measures of driver drowsiness. Virginia Polytechnic Institute & State University, Amsterdam (1985)

[11] Siegmund, K., King, G., Mumford, D.: Correlation of steering behavior with heavy truck driver fatigue. SAE transactions 105(6), 1547–1568 (1996)

[12] Eskandarian, A., Mortazavi, A.: Evaluation of a smart algorithm for commercial vehicle driver drowsiness detection[C]//Intelligent Vehicles Symposium, pp. 553–559. IEEE Press, Istanbul (2007)

[13] Hong, J.E., Min, K.C., Seong-Han, K.: Electroencephalographic study of drowsiness in simulated driving with sleep deprivation. Int. J. of Industrial Ergonomics 35, 307–320

[14] Heckerman, D.: A tutorial on learning with Bayesian networks. In: Jordan, M. (ed.) Learning in Graphical Models, MIT Press, Cambridge (1998)

[15] http://en.wikipedia.org/wiki/Bayesian_network

[16] Bilmes, J.A.: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report, University of Berkeley, ICSI-TR-97-021 (1997)

[17] Moon, T.K.: The expectation-maximation algorithm. IEEE Signal Processing Magazine, 47–70 (November 1996)

Study on Driving Performance of Aged Drivers at the Intersections Compared with Young Drivers

Seunghee Hong[1,2], Byungchan Min[2], and Shun'ichi Doi[1]

[1] 217-20, Hyasi-cho, Takamasu, Kagawa, 761-0301, Japan
[2] 4-218, Deongmyeong-dong, Yuseong-gu, Daejeon, 305-719, Korea
zeele22@naver.com, sdoi@eng.kagawa-u.ac.jp, bcmin@hanbat.ac.kr

Abstract. In the recent aged society, the framework for assisting safe driving should be prepared with understanding the elderly driver's driving performance and their psychological features. The purpose of this study is aimed to obtain the fundamental data of aged driver for their effective assist-system. First, using driving simulator, aged people were observed their driving behaviors in various conditions at intersections compared with young drivers. These behaviors were measured in the condition of right and left turns and crossing. As the results, in particular, significantly slower approaches were observed on every occasion, and the unstable driving behaviors were examined. Next, on the field tests of real running in proving ground, the aged drivers were apt to run rapidly in the case of approaching the crossing compared with young drivers. These driving performances should be interfered with the traffic flow and exposed to the risk of accidents.

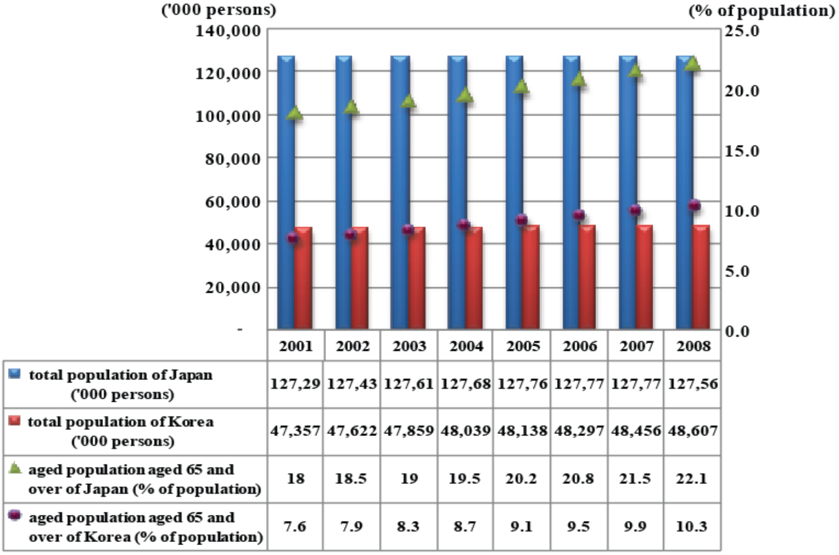
Keywords: Aged Driver, Driving Performance, Intersection.

1 Introduction

Aged is defined as people who are weakened the social role function and self-sustaining function, due to the weakening of physical and physiological and the changes of psychological. Currently, Japan distinguishes a super-aged society, over 22.1% of aged population, and Korea is categorized as an aging society, and is expected to become an aged society by 2019, according to the OECD report in 2010 (see Fig. 1). Therefore, it is required that much more support and assistance. In particular, the most important thing is security, which it is urgent to prepare comprehensive measures that will secure aged people' future. Safe driving assistance would be necessary for one of those things. Recently, aged driver's traffic accidents increased as taking part in various activities. It is a complex and continuous behavior that driving is concerning with the internal and the external situation, and then the behavior quickly and accurately determined (operation) and also to perform repetitive tasks given by approach implicit in a kind of the unconscious memory (implicit memory), according to previous experience. However, the aged who physical function reduced and mental and psychological changed has difficulty while driving.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 156–166, 2011.

Previous studies suggested that they were significant decreases in their ability to split their attention compared with the young driver and rapidly decreased judgments in the event of accidents[1], caused an accident with a high accidental rate per distance traveled, due to they could not appropriate responses[2]. In particular, it is an important issue to be resolved to being involved the fatalities because of weakened physical performance. The development of driver assistance systems for aged drivers to support is urgent. However, appropriate support systems for the aged are few, despite of the development and commercialization. Therefore, it is necessary to characterize the driving behavior and to estimate their internal factors.



	2001	2002	2003	2004	2005	2006	2007	2008
🟦 total population of Japan ('000 persons)	127,29	127,43	127,61	127,68	127,76	127,77	127,77	127,56
🟥 total population of Korea ('000 persons)	47,357	47,622	47,859	48,039	48,138	48,297	48,456	48,607
▲ aged population aged 65 and over of Japan (% of population)	18	18.5	19	19.5	20.2	20.8	21.5	22.1
● aged population aged 65 and over of Korea (% of population)	7.6	7.9	8.3	8.7	9.1	9.5	9.9	10.3

Fig. 1. Total population and Aged population in Korea and Japan

2 Experiment 1

2.1 Methods

Laboratory Environment and Equipment. Driving simulator was designed and display devices were described the frontal scene as same as those of a real vehicle. The information for driving (front, left and right side) provides with LDC 32" monitor attached to the three sides. The environment of the laboratory was controlled. Room temperature was kept 22°and when a simulator was driven, 50dB noise was presented, which was slightly higher than 40dB noise of the common room. (see Fig.2.)

Participants. Subjects were 11 male young drivers, who were 20~30 years old, and 11 aged drivers, who were more than 65 years old. All of them had more than 1 year driving experience. Among them, three old adults dropped out of the test due to Gsimulator sickness, and the data obtained from them were excluded. The average age of participants were 24.2(±3.28), and 73.4(±4.58), respectively, and they had a normal or corrected eyesight with which they had no trouble in perceiving the stimulus presented on the monitor.

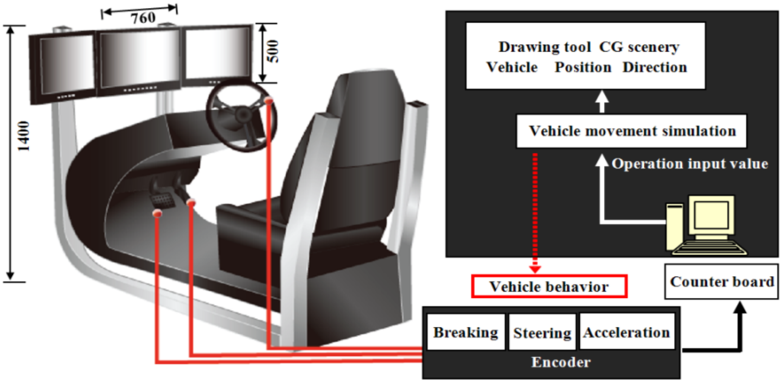
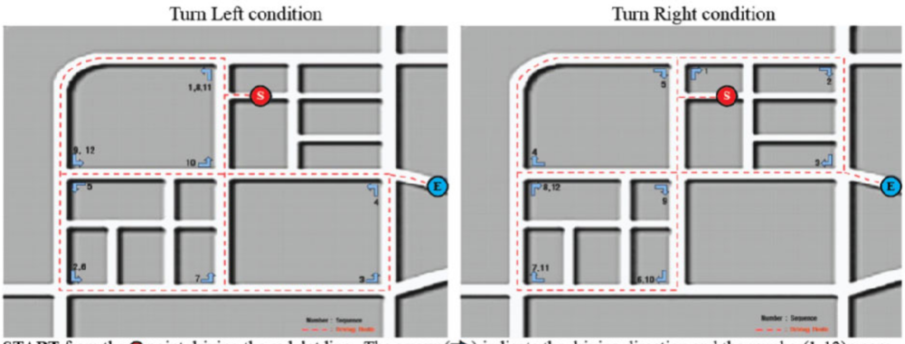


Fig. 2. Driving simulator(GDS-300S, Gridspace Co.)



START from the S point driving the red dot line. The arrows(➦) indicate the driving direction and the number(1-12)means of the turn order. And the driving direction for participants shown on the LCD monitor(Presened before the intersection was 20m) Participants traveled turns (12 times) both Turn Left and Turn Right. Respectively left/right turn intersection are three types.

- Turn Left: T-1(┫) / T-2(┳) / T-3(╋), respectively three times, four times, five times.
- Turn Left: T-1(┣) / T-2(┳) / T-3(╋), respectively three times, four times, five times.

Fig. 3. Driving directions and Order of configurations

Procedure. Before conducting the test, they conducted the practice driving for about 3 minutes in order to adapt to driving simulator. The test consisted of left and right turn conditions of intersection, and the order of the test was counter balanced according to the participants of the test. In 10 min driving scenario which consisted of one - way 2 lanes, the left turn condition was that a driver turned left according to the instruction of direction at the cross intersection and T intersection (No light), and the right turn condition was that a driver turned right according to the instruction of direction like the left turn condition(see Fig. 3.).

Also, by each condition, it consisted of 12 turns, that is, 12 times of conduction. We received data, the accelerator control, brake control, steering from the DS during the experiments. Then to examine the difference of driving according to turn types (left and right turn) at the intersection and age (young and old), 2 * 2 Mixed ANOVA by applying mixed factors design were conducted, in which turn types and age were independent variables and the measures of driving performance were variables.

And then, to examine the difference of driving according to turn types (left and right turn) at the intersection and age (young and old), 2 * 2 Mixed ANOVA by applying mixed factors design were conducted, in which turn types and age were independent variables and the measures of driving performance (approach velocity at the intersection, the passing time at the intersection, the speed when passing the intersection, and handling deviation) were variables.

2.2 Results

Approach Velocity at the Intersection. As a result, there were difference in approach velocity at the intersection, according to turn type and age. In left turn, the main effect of according to age was significant (see Fig. 4) [**p=.000]. But the main effect of according to intersection type (T-1, T-2, T-3) and interaction effect were not significant, respectively [p=.230] and [p=.918]. And in right turn, the main effect according to age [**p=.000] and intersection type [* p=.030] were significant. Also interaction was not significant. Young drivers(Mean of Left Turn =26.194, Right Turn=27.496) approached the intersection faster than aged drivers (Left Turn Mean=17.097, Right Turn=18.442).

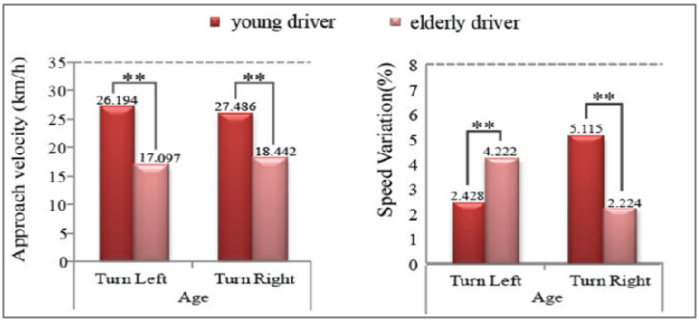


Fig. 4. Approach velocity(left) and Speed variation(right), [*p<.05, **p<.01]

Passing Time. As for passing time, in turn left, the main effect of both age [**p=.000] and intersection types [**p=.000] were significant (see Fig. 5). Also, in turn right, the main effect of both age [**p=.001] and intersection types [**p=.000] were significant. However, in right turn section, the passing time of young drivers was significantly longer than that of aged drivers[**p=.001], In right turn section, the difference of the passing time between two age groups was 1.082s, while in left turn section, that was 4.693s. But all of the interaction effect was not significant. Therefore, the analysis of simple main effect was conducted according to age and turn type. As a result, both left and right turn section were significant, respectively[**p=.001] [**p=.001].

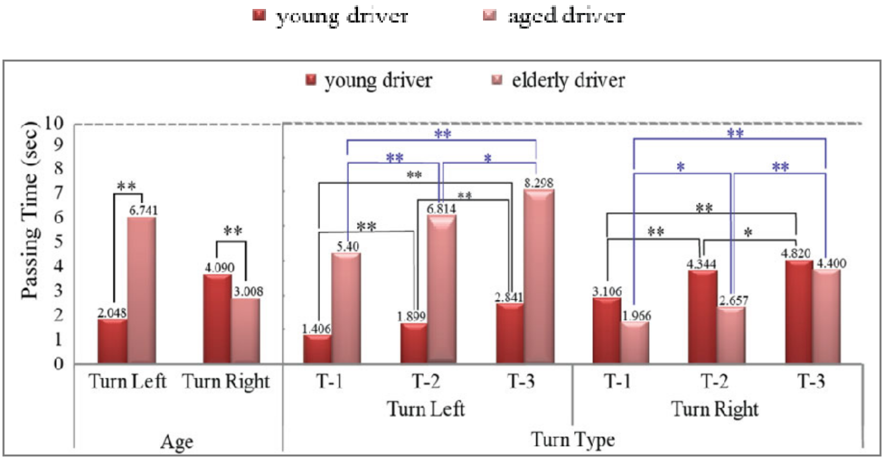


Fig. 5. Passing time thought the intersections (*p<.05, **p<.01).

Speed Variation While Passing Through the Intersection. For speed variation while passing the intersection, as a result of analyzing according to turn types and age, the main effect of both was significant. The speed variation of aged drivers were significantly larger than that of young driver [**p=.004] in turn left section. However, in turn right section, young driver' was larger (see. Fig. 4).

3 Experiment 2

In this experiment, on the field tests of real running in proving ground, the details of aged driver's behaviors were classified according to the process of entering intersections, namely recognizing the crossing and operating the braking. The surroundings of the intersections like stop signs and fences were settled according to experimental conditions in the study. Aged driver's reduced visual acuity was also considered.

3.1 Method

Environment and Equipment. Experiment 2 was conducted with a driver's license center in Kagawa, Japan. Small car (engine bore volume 1,000cc) has been used. The four cameras were installed in the car. These were used to record driving forward scene, changes of the speedometer, driver’s facial scene and a brake pedal behaviors, respectively. Light sensor was installed to detect whether the brake operation.

Participants. Young people who has the usual driving experiance, especially students without problems in 10 participants (men 7, women 3, 22.3-24.0 years, mean age 23 years) and aged people were selected though screening tests (listening survey of personal information, visual acuity, color vision test. MMSE) 10 participants (male 5, female 5, 69-78 years old, mean age 70 years) were take part in the experiment.

Procedure and stimuli. Before conducting the test, they had the practice driving for about 10 minutes in order to adapt to driving ground. Each task was composed 12 times to pass the intersection, namely 4 times trials for three kinds of intersection A,B and C. The experimental time was approximately 20 minutes. Each intersection was composed 6 types according to barrier fences and stop signs (see Fig. 6). The situation of six kinds was provided, according to fences and signs. The fence classified ahead vision, named Easy to view and Difficult to view. And the sign classified to none, general, prominence by installing LED, respectively named No Sign, Typical Sign, and Shining Sign.

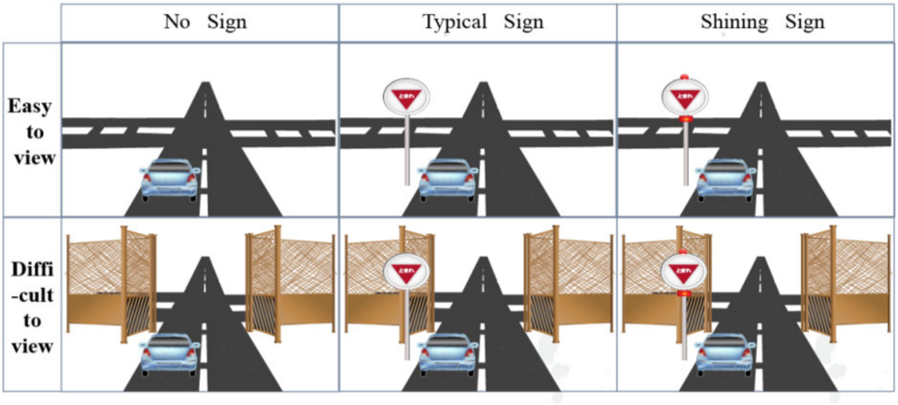


Fig. 6. The six kinds of environment at intersection

These items are measured for every participant in entering the intersections, as followings (see Fig.7).

1. Vo [Km/h]: Velocity at the deceleration onset.
2. Tp[s]: Time from operating brake until of the minimum speed.

3. Lo[m] : Distance from minimum speed position to stop line
4. Jerk [m/s^3]: Rate of acceleration change j from brake operation. a : acceleration v : velocity, r : position

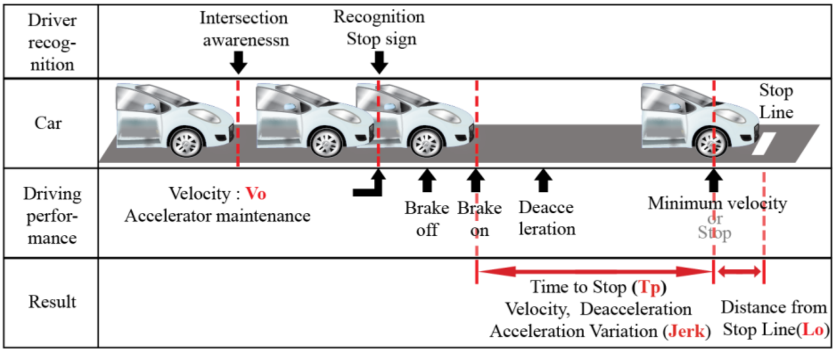


Fig. 7. Measuring of Stop and braking behaviors

3.2 Results

Driving Behaviors While Approach at the Intersection (Fig.8).

- Velocity of the Deceleration onset; Vo[km/h]: Vo of aged driver drivers was faster than that of young drivers in all intersection.
- Time from brake initiation to minimum speed; Tp[s]: Tp of young driver drivers was measured as longer than that of aged drivers in all intersection.
- Distance from minimum speed position to stop line; Lo[m]: Lo of aged driver was farther than that of young drivers' Lo in all intersection.
- Rate of acceleration change from brake operation; Jerk [m/s^3]: Jerk of aged driver was lager then young driver both *Jmin* and *Jmax* in intersection.

Stop According to the Visibility by Stop Sign and Barrier Fence

The variables were classified as four-kinds according to stop behavior and stop consciousness.

- Completely stop: velocity 0
- Willingness to stop: velocity ≠ 0. There was braking.
- Unknown willingness to stop: velocity ≠ 0. There was not both braking and acceleration behavior and affected in the ahead road
- Completely none stop: velocity ≠ 0. There was nothing.

As a result, the percentage of complete stop, willingness to stop, unknown willingness to stop, completely none stop are shown in Table.1. Regardless of age, the completely none stop occupied large percentage in the condition of no sign. Overall, the young drivers increased the rate completely stop compared with aged driver. In addition, the noticeable degree of the conditions sign tended to increase the rate stop regardless of fence. Young drivers were no completely missing where stop sign installed intersections. Fig. 9 shows the percentage of completely stop and none completely stop. Also, completely stop was larger in shining-sign conditions and easy-to view conditions compared with no-sign conditions and difficult-to-view.

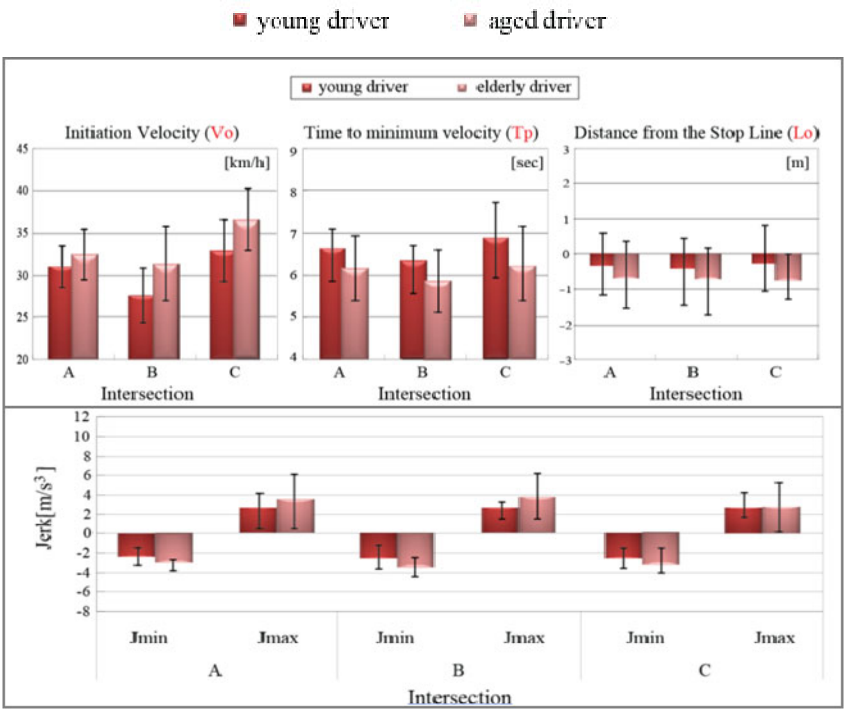
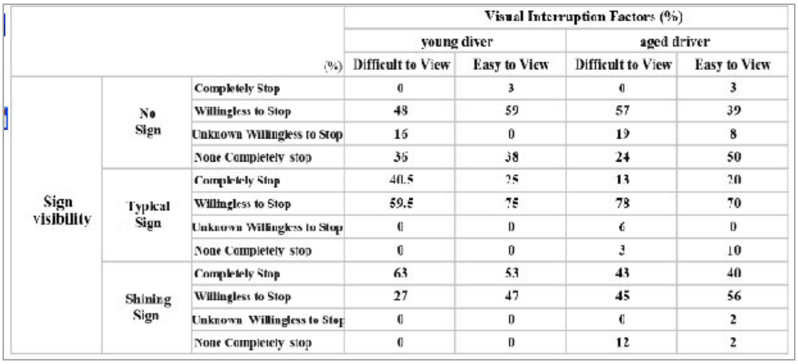


Fig. 8. Comparison of the measured performance (Vo, Tp, Lo and Jerk at the intersection)

Table 1. Percentage according to conditions by sign and fence



			Visual Interruption Factors (%)			
			young diver		aged driver	
		(%)	Difficult to View	Easy to View	Difficult to View	Easy to View
Sign visibility	No Sign	Completely Stop	0	3	0	3
		Willingless to Stop	48	59	57	39
		Unknown Willingless to Stop	16	0	19	8
		None Completely stop	36	38	24	50
	Typical Sign	Completely Stop	40.5	25	13	20
		Willingless to Stop	59.5	75	78	70
		Unknown Willingless to Stop	0	0	6	0
		None Completely stop	0	0	3	10
	Shining Sign	Completely Stop	63	53	43	40
		Willingless to Stop	27	47	45	56
		Unknown Willingless to Stop	0	0	0	2
		None Completely stop	0	0	12	2

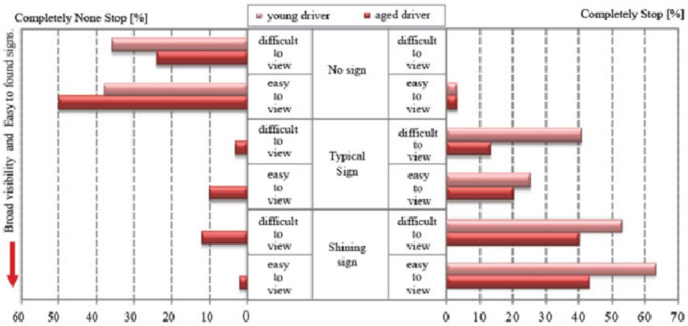
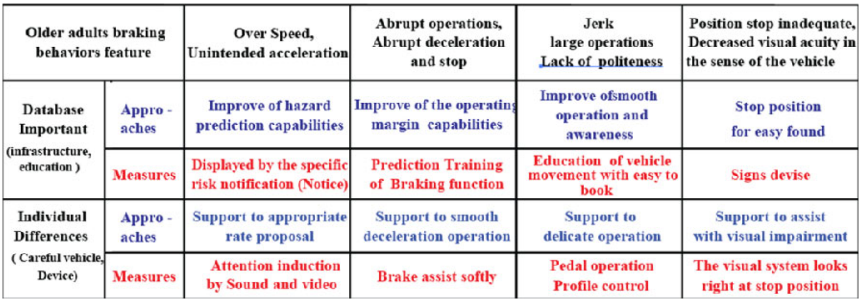


Fig. 9. The rate of completely-stop and none-completely-stop according to age.

Table 2. Database about the aged drivers braking behaviors features



Older adults braking behaviors feature		Over Speed, Unintended acceleration	Abrupt operations, Abrupt deceleration and stop	Jerk large operations Lack of politeness	Position stop inadequate, Decreased visual acuity in the sense of the vehicle
Database Important (infrastructure, education)	Appro - aches	Improve of hazard prediction capabilities	Improve of the operating margin capabilities	Improve ofsmooth operation and awareness	Stop position for easy found
	Measures	Displayed by the specific risk notification (Notice)	Prediction Training of Braking function	Education of vehicle movement with easy to book	Signs devise
Individual Differences (Careful vehicle, Device)	Appro - aches	Support to appropriate rate proposal	Support to smooth deceleration operation	Support to delicate operation	Support to assist with visual impairment
	Measures	Attention induction by Sound and video	Brake assist softly	Pedal operation Profile control	The visual system looks right at stop position

4 Discussion and Conclusion

The purpose of this study was to discuss driving behaviors including turns and straight running at the intersection. Experiments were conducted two times. In Experiment 1, using driving simulator, driving behaviors were measured approaching and passing at intersections without traffic signals based on turn left and turn right classified depending on the type of intersection.

As a result, the aged drivers were slower compared to young driver in the approach velocity at the intersection. However, a significant difference according to the intersection type was not examined. As to the passing time in the intersection, aged drivers were much slower than young drivers, as 6.7 sec and 2.0 sec on turn left. On the other hand, they (4.1 sec) were rather faster compared to 3.0 seconds of young driver. Moreover the simple main effects analysis showed significantly slower both turn left condition and turn right condition at the intersection in orderT-1, T-2, T-3. Next speed variations were measured while through the intersection. The results showed that the variation of the aged drivers were significantly larger in the turn left condition [** p =. 004] and significantly lowers than the young drivers in turn right condition [** p =. 000].

Yeoh Sok Foon (2009) conducted to collect information about driving related practices, knowledge, attitude and confidence of the aged drivers. The results showed that aged drivers who had higher level of self-rated confidence were reflecting higher level of driving ability. However, the response of confidence level score (not confident, somewhat confident, very confident) in many driving situation showed that 83 aged people (20.8%) responded the most difficult not-confident when they do turn right at an intersection without a traffic signal (right-hand; this study indicates left turn, as left-hand) [3]. Aged drives to compensate for the weakening of driving function due to the physical loss, and they show driving behavior avoiding a road where there is much traffic and time zone(rush-hour, etc.) when there is much traffic [4]. In our experiments, Yeoh Sok Foon (2009) the collection research has been experimentally proved, supporting their research. Driving directions before 20m to enter the intersection were indicated by arrows in display the top center, as↰. So, aged drivers accessed significantly slower at the intersection, because they predict the intersection in advance. They showed very large variations while left turn speed approaching at the intersection. It seems unstable the brake pedal operation. So, they found driving at a lower speed, as slow through the intersection. In particular, these driving behavior more definitely has shown according to T1 (each,┫•┣ type) than T2 (┳ type), T3 (╋ type) many more traffic flow, there are a lot of information. That is can be explained aged drivers were slower under a pressure before of approach at the intersection without traffic signal. This hesitant behavior in the intersection rather than prevent traffic flow or interfere with other driver's course, which has greatly the risk of accidents.

Also, we conducted the experiment 2. The actual driving was executed in the course of driver's license training center. Driving directions toward the driver were indicated by the operator of the back seat, so the driver could not predict the location of intersection. In these conditions the direction of running consists only of straight at the intersection. In order to consider visibility and driving attention to the environments, experimental condition was composed several traffic situation by different stop signs and barrier fences. Then, we observed the driving behaviors. As a result, aged drivers traveled much faster than young driver just before detection of the intersections. If they find an intersection they could slow dramatically. They cannot predict the intersection while driving. Then if they detect, the behavior would be a sudden braking. These results can be explained due to the weakness of their vision. They had also a tendency to stop away from the stop line, as a cause of degradation cognitive ability.

In the results by installation of the sign and fence, stop pass percentage was more frequent in the condition of easy-to-view both young and aged drivers. Another completely stop at intersections has increased the rate as easy to detect the signs these results particularly were noticeable from the aged. Their speed was very fast, 40km/h, in the condition of easy-to-view when through the intersection. This can be explained that aged drivers drive careless and determined and sure driving ability. But driving around the environment (time of day, traffic flow, etc.), there are very many factors. Their decision is likely to be an error, as the degenerations of a dynamic vision and the distance vision to measure in front of the object, cognitive ability, also physical reflexes backwards or stuff in their ability to respond to unexpected situations. That led to a fatal accident.

In these experiments, we discussed the variable about the driving performance of comparing aged drivers and young drivers when approach and pass the intersection. The results show that aged driver were very confident in driving and very careless driving when they are convinced the situation of ahead as visually. However, they bring to pressure at the left turn showed the slow and hesitant driving behavior, which prevent the traffic flow and can cause confusion for other drivers. In experiment 2 suggests that they given some of driving stimulus (support) could be improved, based on increasing the rates of stop according to the condition on sign. So, it is urgently needed considering the aged characteristics. According to these results, the basic database of the aged driving characteristics was verified to establish the effective driving support system.

In this study, there were limitations to be conducted in that the DS experiments in South Korea (Left Handle) and the real driving test in Japan (right handle). But the results is considered that would be used as a very important to study the aged support system in the future.

References

1. Kahneman, D., Ben-lshai, R., Lotan, M.: Relation of a test of attention to road accidents. Appilied Psychology 58, 113–115 (1973)
2. Anderson, J.R., Lebiere, C.: The atomic components of thought, p. 490. Lawrence Erlbaur Assiciates, Hillsdale(1998)
3. Foon, Y.S.: Driving practices of older Malaysian drivers: The influence of Knowledge. Attitude and confidence (2009)
4. Lee, S.C.: Lee, Psychological effects on aged driver's traffic accidents. Korean Journal of Psychological and Social Issues 12, 149–167 (2006)

The Influence of False and Missing Alarms of Safety System on Drivers' Risk-Taking Behavior

Takayuki Masuda[1], Shigeru Haga[2], Azusa Aoyama[2], Hiroki Takahashi[2], and Gaku Naito[2]

[1] Human Science Division, Railway Technical Research Institute,
2-8-38 Hikari-cho, Kokubunji-shi Tokyo, 185-8540, Japan
[2] School of Contemporary Psychology, Rikkyo University,
1-2-26 Kitano, Niiza-shi, Saitama, 352-8558, Japan
masuda@rtri.or.jp, haga@rikkyo.ac.jp,
foreverkizamikakao@gmail.com, spellmiss_hi6i@auone.jp,
10um007x@rikkyo.ac.jp

Abstract. This study investigates the influence of false and missing alarms of safety system on drivers' risk-taking behavior by laboratory experiments. The task is to move a vehicle from below to top through an intersection displayed on a PC monitor without colliding with crossing traffic. Participants performed the task under different experimental conditions with different types of system failure: (1) no failure, (2) false alarm, (3) missing alarm, and (4) no information. We conducted two experiments. The difference between Experiment 1 (E1) and Experiment 2 (E2) is the frequency of false or missing alarms: erroneous alarms occurred twice as many in E2 as E1. The differences of the result between E1 and E2 indicate that the different frequencies of missing alarm have a different effect on risk-taking behavior.

Keywords: negative adaptation, risk-taking, system failure.

1 Introduction

To date, remarkable progress in vehicle technology has made it possible to introduce various driving support systems, such as ACC (adaptive cruise control), ISA (intelligent speed adaptation), AAP (active accelerator pedal) and VES (vision enhance system) to the market. These systems contribute to the safety improvements; however, negative adaptation may spoil the expected safety effect of these systems.

The negative adaptation means undesirable behavioral changes, which may occur following the introduction of safety measures such as driving support systems[1]. There are ample evidences of the occurrence of negative adaptation[2], [3], [4], though it does not always occur. What is important is to identify factors affecting negative adaptation and to find the way to mitigate the negative effect caused by the negative adaptation.

Various psychological factors relate to negative adaptation. Trust to the safety system is one of the most important psychological factors affecting the negative adaptation. Over-trust may lead to misuse and distrust may lead to disuse[5]. Both of

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 167–175, 2011.

them impair safety. Misuse refers to the problems that occur when people rely on the system and use it inappropriately. Disuse refers to the problem that occurs when people reject to use the system.

Trust is affected by false and missing alarms. Among the studies conducted on the trust and false and missing alarms, some studies have indicated false and missing alarms of safety systems lead delay of response to alarms due to over-trust[6]: however, there is few studies focused on the relation between false and missing alarms of safety systems and risk-taking behavior. This study investigates the influence of false and missing alarms of safety system on drivers' risk-taking behavior by laboratory experiments.

The purpose this study was to investigate the influence of false and missing alarms of safety system on drivers' risk-taking behavior. We conduct two experiments.

2 Experiment 1

2.1 Procedure

Participants. Eleven people (six male, six female) participated in the study; they had the mean age of 20.64 years and the mean driving experience of 1.55 years. Ethical permission was granted by the Department of Psychology at the Rikkyo University. All participants were aware of their right to withdraw from the study at any time and had a full debriefing about the aims of the study.

Equipment. We collected data using the experiment software (developed with Microsoft Visual Basic 2005). Experiment software was installed to the PC (Dell XPS720) and controlled with a USB device (Microsoft Side Winder Joystick) connected to the PC. Output was displayed on the monitor (Dell SE197FPS) at 1024 × 768 pix.

Task. The task was to move a vehicle from below to top through an intersection displayed on the PC monitor without colliding with crossing traffic (Fig. 1). Participants moved the joystick to the left or right to “look” at approaching traffic, otherwise crossing traffic could not be visible. Participants performed the task under different experimental conditions with different types of system failure: (1) no failure, (2) false alarm, (3) missing alarm, and (4) no information.

The following illustrates the system providing information an approaching traffic. The system has four lights indicating the existence of traffic on four traffic lanes as shown on the monitor. If the system detects an approaching vehicle as far as two vehicle lights out of the visible range from the intersection.

The experimental task consisted of four “blocks”. It took about one hour to go through one block per person. Intervals between blocks were more than two hours. One block consisted of four sessions. Participants go through “no information” condition in Block 1, them "no failure" condition. The order of "false alarm" and "missing alarm" condition was counter balanced (Table 1). One block consisted of four sessions. In each block, the first session was a pilot session. One session consisted of six sections. Participants took two minutes' rest between sessions. The speed of approaching vehicle and its number in one section is as shown in Table 2.

In each experimental condition, "events" occurred three times (Fig. 2). In "no information", "no failure" and "missing alarm" conditions, one vehicle approached the intersection on each traffic lane and its speed was selected from 16, 17 or 18 pix/100 msec. In missing alarm condition, light did not turn on during the event. In "false alarm" condition, the information was provided at the same timing as if a vehicle was approaching at the speed selected from the same range under other conditions as referred in Table 3.

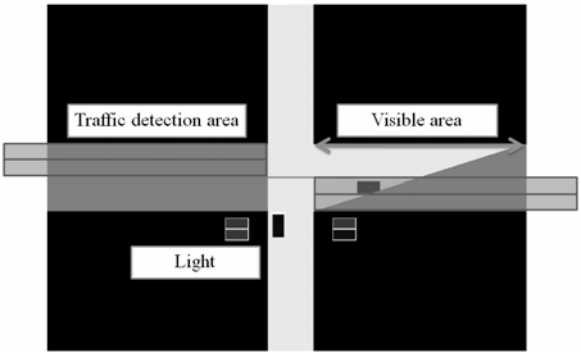


Fig. 1. Experimental task: Participants can "look" left or right with a joystick

Table 1. Experimental design

Block 1				Block 2				Block 3				Block 4			
No information				No failure				False/missing alarm				False/missing alarm			
trial	S1	S2	S3	trial	S1	S2	S3	trial	S1	S2	S3	trial	S1	S2	S3

Table 2. Speed of approaching vehicle and its number on each traffic lane on one section

Speed (pix/100msec)	Number
16	2
17	2
18	2
31	1
32	1

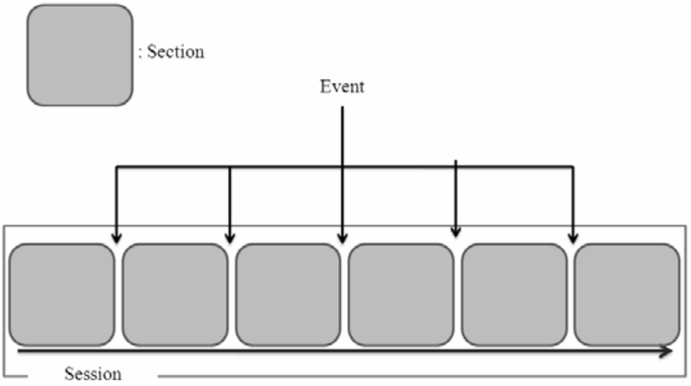


Fig. 2. Events occur in three out of five randomly selected intervals between sections

Table 3. State of vehicle and information in event of

	Vehicle	Information
No information	Approaching	None
No failure	Approaching	Provided
False alarm	No	Provided
Missing alarm	Approaching	Provided

2.2 Results

Effect of Types of System Failure. Analysis of variance (ANOVA) was conducted to examine the effects of different types of system failure on risk-taking behavior. Dependent variables were number of intersection crossings in one block, number of looking left or right per crossing and number of collisions in one block. Number of intersection crossings "in one block" means the total number of intersection crossings in three sessions except the trial session in each experimental condition. Number of collisions in one block was calculated in the same way. The results indicated that the main effect of the types of system failure on the number of intersection crossings was significant ($F(3,30) = 8.10$, $p < 0.01$). The LSD multiple comparison indicated that the number of intersection crossings in no failure, missing alarm and false alarm conditions was significantly more than that in no information condition (Fig. 3). The results showed that the main effect of the types of system failure on the number of looking left or right was significant ($F(3,30) = 8.02$, $p < 0.01$). The LSD multiple comparison indicated that the number of looking left or right in no failure, missing alarm and false alarm conditions was significantly fewer than that in no information condition (Fig. 4). The results showed that the main effect of the types of system failure on the number of collision was significant ($F(3,30) = 26.80$, $p < 0.01$). The LSD multiple comparison indicated that the number of collision in no failure, missing alarm and false alarm conditions was less than that in no information condition. The number of collision in missing alarm was significantly more than that in no failure condition (Fig. 5).

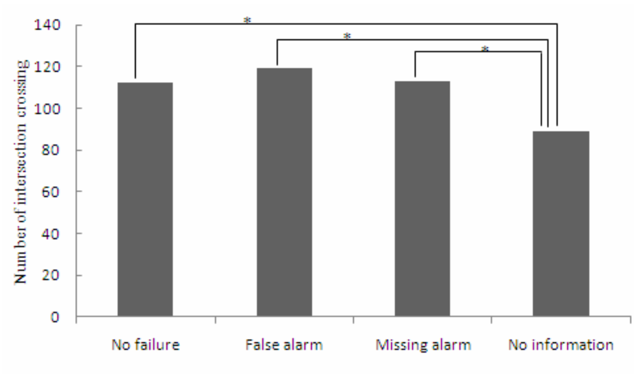


Fig. 3. Number of intersection crossings in one block (*: *p*<.05)

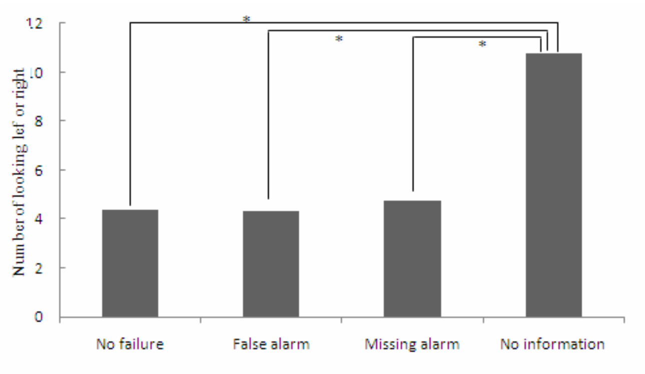


Fig. 4. Number of looking left or right per crossing (*: *p*<.05)

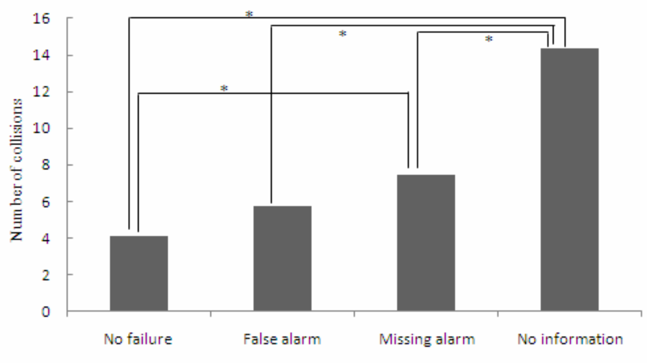


Fig. 5. Number of collisions in one block (*: *p*<.05)

2.3 Discussion

Effect of Types of System Failure. The purpose of this experiment is to investigate the effect of the system failure on the risk taking behavior. As shown above, the effect of system failure on risk-taking behavior was not seen. There was a difference in the number of intersection crossings and looking left or right between no information condition and other conditions, but not false alarm and missing alarm conditions. This is because of the lack of the frequencies of false and missing alarms and could not affect on behaviors. Participants evaluated that the frequencies of system failure was low (false alarm: $M = 1.73$, missing alarm: $M = 1.73$, five-scale questions).

The result showed that participants collided more in the missing alarm condition than in the no-failure condition. However, the collisions occurred due to missing alarms was few (twice among all participants). In addition, there were no differences in the number of looking left or right and the number of intersection crossings among conditions. This result may indicate that participants were confused by missing alarm, and could not judge the timing to cross the intersection. The difference of the effect among types of system failures may be seen if relative frequency of missing or false alarms was higher.

We conducted Experiment 2. The difference between Experiment 1 (E1) and Experiment 2 (E2) was the frequency of false or missing alarms: erroneous alarms occurred twice as many in E2 as E1.

3 Experiment 2

3.1 Procedure

Participants. Sixteen people (three male, thirteen female) participated in the study; they had a mean age of 22.25 years and mean driving experience of 2.93 years. Ethical permission was granted by the Department of Psychology at the Rikkyo University. All participants were aware of their right to withdraw from the study at any time and had a full debriefing about the aims of the study.

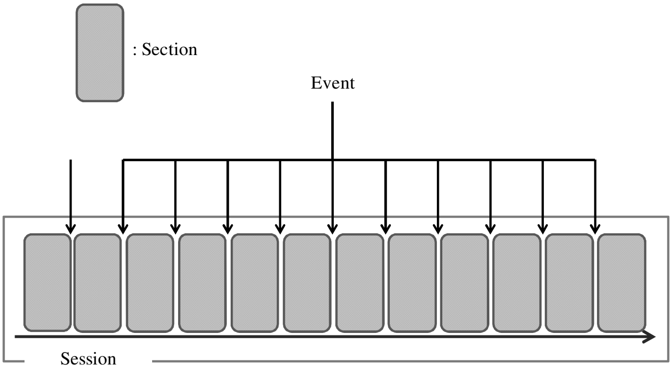


Fig. 6 Event occurs in 6-7 out of eleven randomly selected intervals between sections, with the first intervals always having the event

Equipment and Task. In E2, the same equipment and task was used. The difference between Experiment 1 (E1) and Experiment 2 (E2) was the frequency of false or missing alarms: erroneous alarms occurred twice as many in E2 as in E1. Events occur in 6-7 out of eleven randomly selected intervals between sections, with the first interval always having the event.

Table 4. Speed of approaching vehicle and its number on each traffic lane in one section

Speed (pix/100msec)	Number
16	2
17	2
18	2
31	1

3.2 Results

Effect of Types of System Failure. Analysis of variance (ANOVA) was conducted to examine the effects of different types of system failure on risk-taking behavior. The results showed that the main effect of the types of system failure on the number of intersection crossings was significant (*F*(3,45) = 12.09, *p* <0.01). The LSD multiple comparison indicated that the number of intersection crossings in no information condition was significantly fewer than that in no failure condition, missing alarm and false alarms conditions, and that the number of intersection crossings in missing alarm condition was significantly fewer than that in false alarm condition (Fig. 7). The results showed that the main effect of the types of system failure on the number of looking left or right was significant (*F*(3,30) = 8.50, *p* <0.01). The LSD multiple comparison indicated that the number of looking left or right in no failure, missing alarm and false alarms conditions was significantly fewer than that in no information condition (Fig. 8). The results showed that the main effect of the types of system failure on the number of collision was significant (*F*(3,30) = 28.49, *p* <0.01). The LSD multiple comparison indicates that the number of collision in no failure, missing alarm and false alarms conditions was significantly less than that in no information condition. The number of collision in missing alarm was more than that in no failure condition (Fig. 9).

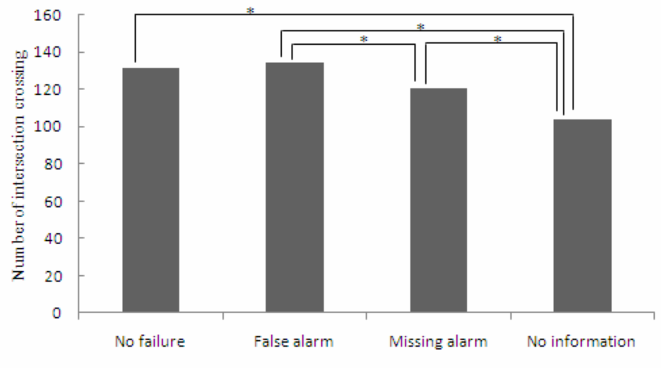


Fig. 7. The number of intersection crossings in one block (*: *p*<.05)

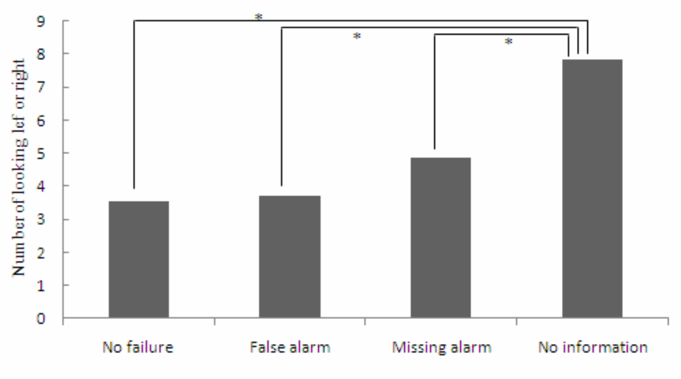


Fig. 8. The number of looking left or right per crossing (*: *p*<.05)

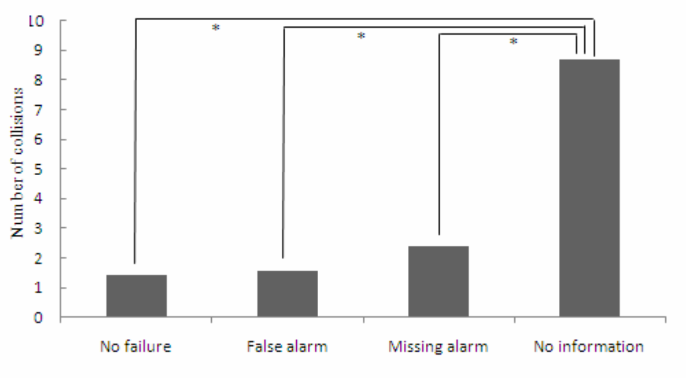


Fig. 9. The number of collisions in one block (*: *p*<.05)

3.3 Discussion

The result indicated that missing alarm affected risk-taking behavior. In the no-failure condition and false-alarm and missing-alarm conditions, drivers attempted to cross more than in no-information condition. In the missing-alarm condition, drivers attempted to cross more than in false condition. These results suggest that missing alarm possibly suppress the risk-taking behavior. However, there was no difference in the number of collisions between false alarm and missing alarm conditions. This result does not mean that missing alarm is safer than false alarm.

4 General Discussion

We conducted two experiments to examine the effect of system failure on risk-taking behavior. The result of E1 did not show the difference under the effect of type of system failure on risk taking behavior. The result of E2 indicated that missing alarm affected risk-taking behavior. The differences of the result between E1 and E2 indicated that the different frequencies of missing and false alarms have a different effect on risk-taking behavior.

Although this study demonstrated the difference of effect of the type of system failure on risk-taking behavior, we need the following further researches.

We need to investigate a long-term effect of system failure. For examples, risk-taking behavior will increase longitudinally even if drivers use the systems effectively when system failures rarely occur. We need to conduct the experiment on conditions that missing and false alarms occur in various frequencies. We also need to investigate the case that has both missing and false alarms occur in one block. In reality, one driver may experience several types of system failures.

Understanding the relationship between types of system failure, its frequencies and behavioral changes may make it possible to help system design (e.g., the criteria for deciding tolerable frequencies of system failures).

References

1. OECD, Behavioural adaptations to changes in the road transport system: report / prepared by an OECD scientific expert group. Paris. Organisation for Economic Co-Operation and Development, Paris Washington, D.C (1990)
2. Stanton, N.A., Pinto, M.: Behavioural compensation by drivers of a simulator when using a vision enhancement system. Ergonomics 43, 1359–1370 (2000)
3. Hoedemaeker, M., Brookhuis, K.A.: Behavioral adaptation to driving with an adaptive cruise control (ACC). Transportation Research Part F: Traffic Psychology and Behaviour 1, 95–106 (1998)
4. Rudin-Brown, C.M., Parker, H.A.: Behavioural adaptation to adaptive cruise control (ACC): Implications for preventive strategies. Transportation Research Part F: Traffic Psychology and Behaviour 7, 59–76 (2004)
5. Lee, J.D., See, K.A.: Trust in Automation: Designing for Appropriate Reliance. Human Factors 46, 50–80 (2004)
6. Abe, G., Richardson, J.: The influence of alarm timing on driver response to collision warning systems following system failure. Behaviour & Information Technology 25, 443–452 (2006)

Estimation of Driver's Arousal State Using Multi-dimensional Physiological Indices

Mieko Ohsuga[1], Yoshiyuki Kamakura[2], Yumiko Inoue[2], Yoshihiro Noguchi[3],
Kenji Shimada[3], and Masami Mishiro[4]

[1] Faculty of Engineering, Osaka Institute of Technology. 5-16-1 Omiya, Asahi-ku, Osaka-city,
Osaka 535-8585, Japan
[2] Faculty of Information Science and Technology, Osaka Institute of Technology. 1-79-1
Kitayama, Hirakata-city, Osaka 573-0196, Japan
[3] National Institute of Advanced Industrial Science and Technology (AIST). 1-1-1, Umezono,
Tsukuba-city, Ibaraki, 305-8568, Japan
[4] Information Technology Laboratory, Asahi Kasei Corporation. 3050 Okada, Atsugi-city,
Kanagawa 243-0021, Japan
{ohsuga@bme,kamakura@is,yumiko@is}.oit.ac.jp,
{y.noguch,shimada.kb}@aist.go.jp,
mishiro.mc@om.asahi-kasei.co.jp

Abstract. The goal of our research is to develop a method to assess the arousal states using facial images of drivers. Multi-dimensional physiological indices are expected to be alternative external criteria of arousal states to manual coding of facial expression which require a lot of human resources. Changes in multi dimensional physiological indices (i.e., blink categories, skin conductance, EEG alpha wave, respiration, heart rate variability) depending on the arousal states defined by the combination of "arousal level" and the presence of "effort" to wake up were studied. Multiple linear regression analysis was also executed using one of the face scores ("arousal level" or "effort") as dependent variables and the physiological indices as explanatory variables. Relatively high multiple correlation coefficients were obtained, however, the number and combination of selected indices showed great differences between individuals. To obtain common equations is an issue in future.

Keywords: driver behavior, arousal level, drowsiness, facial expression, blinks, electro-oculogram, skin conductance, electroencephalogram, respiration, heart rate, heart rate variability.

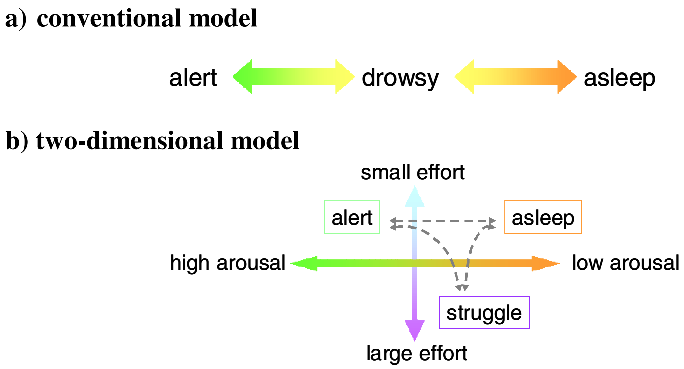
1 Introduction

The assessment of driver's arousal levels is one of the most important issues for the development of adaptive driving support systems. Despite a long history of research in this field [1-7], the method to efficiently assess the arousal levels has not been realized. Particularly in the case of driving situation, the transition of the arousal levels does not simply change monotonically over time. The situation is rather more complex since the drivers tend to be compelled to overcome his/her drowsiness states.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 176–185, 2011.

Physiological indices of the drivers are consequently expected to show complex variations. The effective strategy to prevent accidents caused by drowsy driving is supposed to depend on whether the driver is aware of his/her lowering arousal state and whether he/she is making effort to wake up.

From this point of view, we introduced a two-dimensional model of arousal state of which axes are arousal level and the amount of efforts to keep arousal level [8]. Figure 1 shows the conventional model (a) and the proposed two-dimensional model (b) for transition of arousal states. In addition we proposed new criteria for manual rating of facial expressions to characterize the battle against drowsiness [8].



a) conventional model

b) two-dimensional model

Fig. 1. Two-dimensional model of arousal states compared to the conventional model

The goal of this research is to develop a method to assess the arousal states by image processing of facial images during driving. We already reported the possibility of driver-independent assessment of arousal states from video sequences using HMMs (Hidden Markov Models) [9]. A large amount of dataset is required to refine a machine discriminator by machine learning. As manual rating of facial expressions is time-consuming task, we expect to estimate arousal states using multi-dimensional physiological measures which can be obtained continuously without human hand. During a real-life driving, it is difficult to measure physiological responses without putting an extra burden on the driver, however, it is not a critical problem in the experimental setting. In the present paper, the changes in multi-dimensional physiological measures are referred to the changes in arousal state scores obtained by manual rating of facial expressions.

2 Methods

2.1 Data Collection

The experimental data we reported at HCII2007 [10] were examined again.

Participants. Fifty-nine paid volunteers who gave written informed consent by their free will participated in that experiment. Participants were all licensed drivers but their driving experiences varied.

Experimental Task and Recordings. Each participant was instructed to execute a simulated driving task for fifty minutes. A simulated driving environment was set up by using a commercial video game system projected on a 100-inch screen in front of the participant. A monotonous driving course, a circular circuit with little curve and no turn, was created for this experiment. A game controller with a steering wheel, an accelerator and a brake pedal was used for driving operation. The participants were required to respond as quickly as possible, by pressing a small button on the steering wheel, when a circle displayed on a 7-inch LCD became larger. Facial images of participants were captured by two CCD cameras and video recorded. Figure 2 shows the experimental settings.

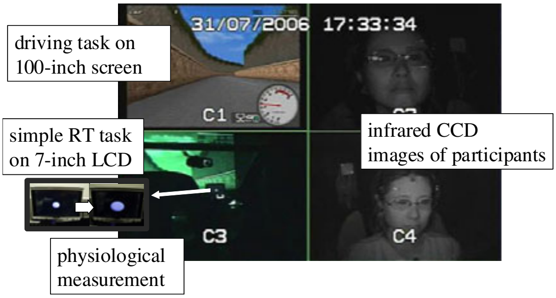

31/07/2006 17:33:34
driving task on 100-inch screen
C1
simple RT task on 7-inch LCD
infrared CCD images of participants
C3
C4
physiological measurement

Fig. 2. Experimental settings and participant's view of driving task

Physiological Measurement. Multi-dimensional physiological recordings were carried out. These include the vertical and horizontal electro-oculogram (EOG), occipital midline electroencephalogram (EEG) recorded by using linked-ear references, skin conductance levels and responses (SCL, SCR), chest electrocardiogram (ECG) and respiration. The Ag/AgCl electrodes were used for EOG, EEG, SC, and disposable electrodes were used for ECG. Respiratory movement was measured using a carbon tube sensor attached around abdomen. These data were amplified and sent to a PC by a digital multi-channel amplifier for biological use (Polymate AP1132, TEAC). The time constants of the amplifier were set to 0.3 s for EEG, and 3.0 s for EOG and respiration, respectively. Skin conductance between two finger-tips was converted to voltage by an EDA unit (AP-U030) designed based on a circuit recommended by Society for Psychological Research. The original sampling rate was 1 kHz, however physiological data was analyzed after re-sampled to 200 Hz.

Procedure. The first part of the experiment was a 2-minute session where the participant was asked to be seated at the driving simulator and look at some fix points on the front screen with no restriction regarding their eye-blinks. The data recorded in this part were used for calibration purpose. A short training session was allowed so that the participant became used to the driving operation. The second part of the experiment was the actual 50-minute driving task with simple reaction time task. The experiment was aborted at anytime when the participant complained about some uncomfortable symptoms such as motion sickness.

2.2 Quantification Procedure of Indices

Facial Scores. Three trained experimenters rated driver's faces from two viewpoints based on the proposed two-dimensional model. One is the arousal level; 1 (high arousal) to 4 (low arousal) and the other is the presence of effort (α: none, β: effort) to wake up. The recorded video was divided in 20-second segments which were presented to raters in a random order. After rating, the scores were ordered in time and averaged for every 1 minute, i.e. nine scores (3 scores for every 20-second by 3 raters) were used for the calculation.

Physiological Measures. The average of each physiological measure obtained as follows was also quantified for every 1 minute.

By applying the FFT spectral analysis, the average amplitudes of EEG components, i.e., theta (4-8 Hz), alpha (8-13 Hz) and beta (13-30 Hz) were obtained, The EEG alpha index *(EEG-αI)* was defined as the ratio of the amplitude of alpha component to the sum of those of three components.

Instantaneous heart rate was calculated from R-R intervals on ECG and the trend in heart rate *(HR)* was extracted by low-pass filter with a cut-off frequency of 0.04 Hz. Mid-frequency component of heart rate variability *(HRV-MF)* and high-frequency component *(HRV-HF)* were obtained using band-pass filters with 0.08-0.12 Hz and 0.12-0.5 Hz, respectively.

Skin conductance level *(SCL)* was extracted from skin conductance *(SC)* by low-pass filter with a cut-off frequency of 0.04 Hz. Skin conductance responses *(SCR)* was obtained subtracting *SCL* from *SC* and the averaged absolute *SCR* was calculated.

The average amplitudes of respiration for all frequency band *(Resp-all)* and for mid frequency band ranged 0.08-0.12 Hz *(Resp-mf)* were obtained by the FFT analysis. Peak frequency of respiration *(Resp-pf)* and amplitude weighted frequency of that *(Resp-gf)* were also calculated. The absolute difference between *Resp-pf* and *Resp-gf* *(Resp-diff)* was introduced as a measure of respiratory instability.

Blink Categories. The procedure to find blinks in vertical EOG and to derive three characteristic parameters from each blink wave was mentioned in the previous report [10]. The parameters are the peak height (blink amplitude; *PA*), the rising time (closing duration; *T1*) and the falling time (opening duration, *T2*) shown in Fig.3. In the previous work, the classification of all blinks during the simulated driving of each participant was performed by K-means clustering and the obtained classes were classified again into five blink categories inspecting not only the combination of characteristic parameters but also the temporal changes of occurrence frequency (see Table.1). In this paper, the assignment of the blink categories was done for individual blink not for the class of blinks based on the average parameters *PA*, *T1* and *T2* of the blink category '*A*'. Each parameter of the blink was compared with the average parameter of the standard blink category '*A*' and regarded as 'standard' when it fell within the ranges ±5% for *PA* and ±10% for *T1* and *T2*, whereas 'high' and 'low' indicated the levels above and below these ranges, respectively. However the results of individual assignment by these two methods were different, the temporal changes in occurrence frequency of each category were quite similar. The number of blinks which classified into each category was counted for every 1 minute and the ratio of the total number of blinks during the same period was also calculated.

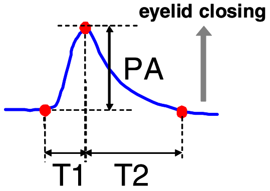
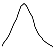
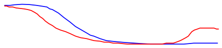
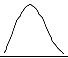
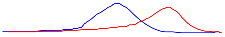
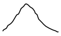
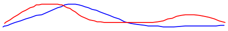
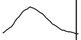
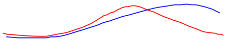
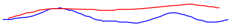


Fig. 3. Blink parameters extracted from v-EOG [8]

Table 1. Blink categories and the typical temporal changes in the number of blinks classified into each category (revised from [8])

blink categories			typical temporal changes
A		standard	
B		large PA or longT1	
C		small PA	
D		not large PA & long (T1 and/or T2)	
E		small PA & short (T1 & T2)	

2.3 Statistical Analysis

ANOVA. Effects of arousal states on physiological measures and distribution of blink categories were studied using one-way ANOVA with the significant level α=0.05. Arousal states were defined by the combination of arousal level 1 (alert) ~ 3 (low arousal) and the presence of effort to wake up (α: no effort, β: effort). Arousal levels were decided by the averaged face score on arousal level (s1) as follows; 1: s1<1.33, 2: 1.33 <=s1<2.5, 3: s1>=2.5. The presence of effort was decided by the averaged face score on effort (s2) as follows; α: s2<0.33, β: s2>=0.33.

Multiple Linear Regression Analysis. Multiple linear regression analysis was also executed using one of the face scores (arousal level or effort) as dependent variables and the physiological indices as explanatory variables by step-wise methods (addition of variable: p<=0.05, elimination of variable: p>=0.15) for each participant.

3 Results and Discussions

3.1 Changes in Arousal States

Seven out of fifty-nine participants aborted the experiments due to their uncomfortable symptoms resembled motion sickness. Video recorded face images could not rated for nineteen participants. The major cause was a partial protruding of face image from the camera frame. Out of remaining thirty-three participants, one was omitted by a partial lack of physiological data and the other four were excluded because of too small number of blinks. In consequence, twenty-eight participants were included for the further analysis. Table 2 shows the grouping of participants depending on the variation width of arousal states.

Table 2. The grouping of participants depending on the variation width of arousal states

group	Arousal states					N		
	1α	2α	2β	3α	3β			
5	○	○	○	○	○	5	12	28
4a	○	○	○	×	○	5		
4b	×	○	○	○	○	1		
4c	○	○	○	○	×	1		
3	○	○	○	×	×	11	16	
2	○	○	×	×	×	3		
1	○	×	×	×	×	2		
N	27	25	23	7	11	28		

No participant showed the arousal state ‘1β’ (‘high arousal’ and ‘effort’), that is reasonable from the definition. Only five participants showed other five arousal states. Seven participants were lacking another state. Remaining sixteen participants did not reach arousal level ‘3’. Three of them showed only ‘1α’ and ‘2α’ and no ‘effort’. Two showed no change in arousal state.

3.2 Characteristic Changes in Physiological Measures Depending on Arousal States

The data from twelve participants who reached arousal level ‘3’ belonging to the groups ‘5’,’4a’,’4b’ and ’4c’ in Table 2 were analyzed. The representative values for every 1-min period were standardized within participants for the existing arousal states (four or five states for each participant). The period with less than five blinks was not included.

Figure 4 and 5 show the mean values and standard deviations of the indices for each arousal state with the results of one-way ANOVA.

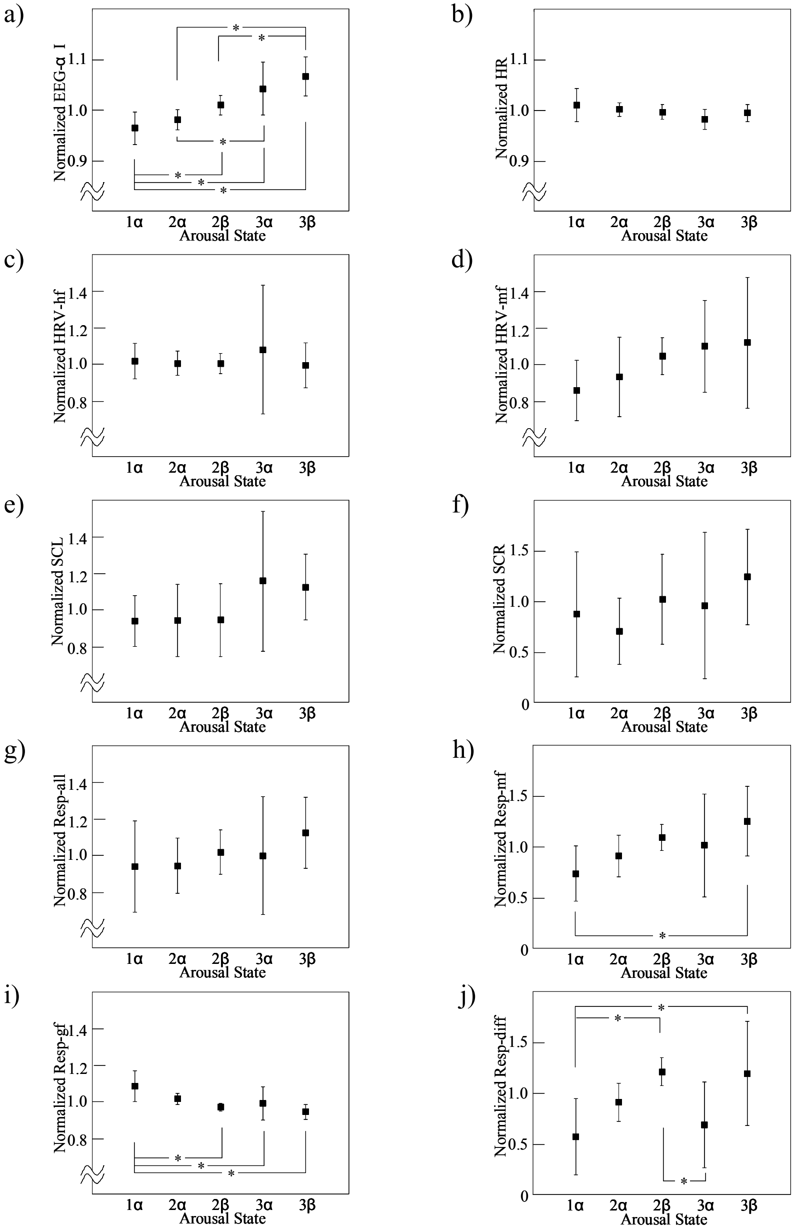


Fig. 4. Changes in physiological indices depending on arousal states: a) EEG-αI, b) HR, c) HRV-hf, d) HRV-mf, e) SCL, f) SCR, g) Resp-all, h) Resp-mf, i) Resp-gF, j) Resp-diff

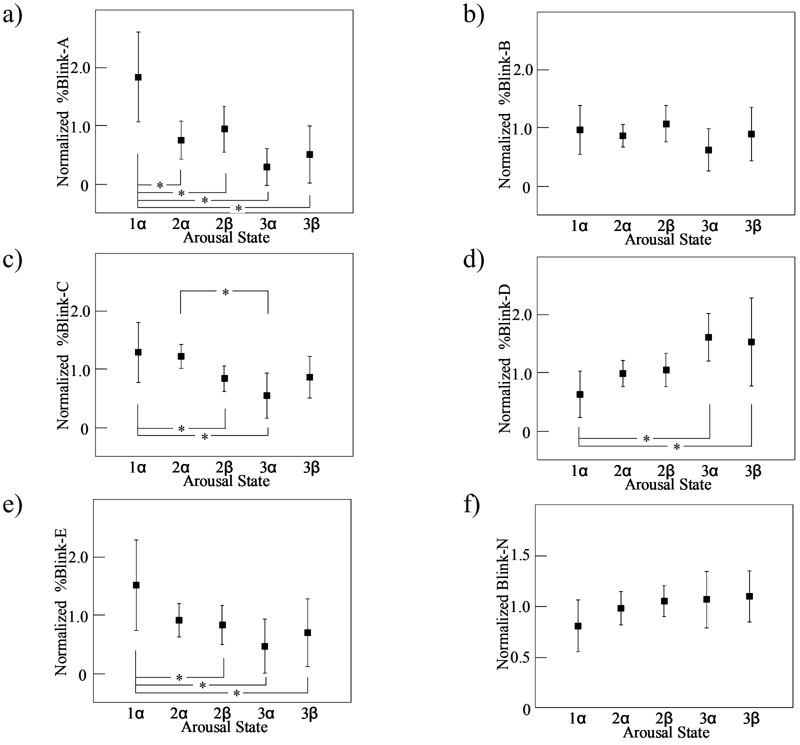


Fig. 5. Changes in blink indices depending on arousal states: a) %Blink-A, b) %Blink-B, c) %Blink-C, d)%Blink-D,e) %Blink-E, f) Blink-N

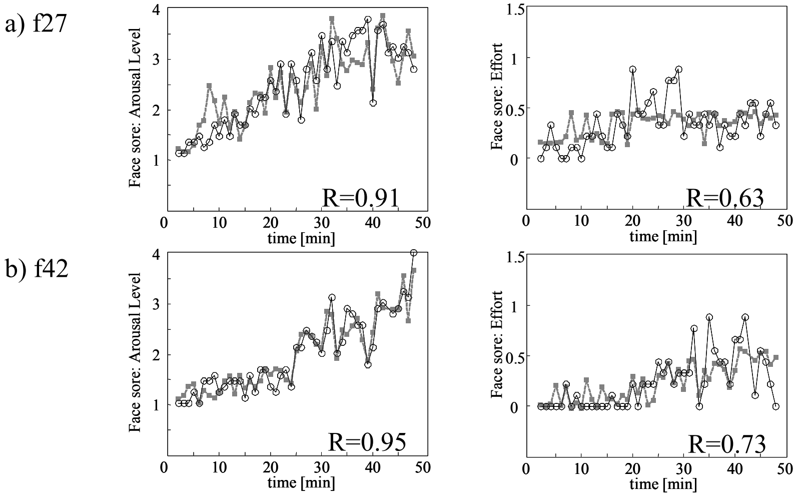


Fig. 6. Real (-o-) and estimated (--■--) face scores using multiple linear regression model. left: 'arousal level', right: 'effort'

EEG-αI increased with the progress in changes in arousal states. Significant differences were observed in six pairs of states. *HR* tends to decelerate and *HRV-hf* tends to increase at '3α' reflecting parasympathetic dominance. However, these changes are not significant because of a large variance. *HRV-mf* increased with the progress in changes in arousal states, but the individual differences are large which possibly depends on the differences in respiration changes. *SCL* elevation in arousal state '3β' might be the effect of 'effort', the elevation in mean value of *SCL* and a large standard deviation in '3α' suggests the contamination of '3α' with '3β'. *SCR* tends to be larger at '2β' and '3β' than '2α' and '3α' suggesting the effect of by 'effort'. However, neither change in SC measures was significant. Some respiratory measures showed characteristic changes depending on arousal states. *Resp-gf* is significantly higher at '1α' than '2β', '3α' and '3β', while *Resp-mf* is smaller at '1α' than '3β'. Although not significant, a higher *Resp-gf* and a smaller *Resp-mf* are observed at '3α'. *Resp-diff* showed a significant increase at '2β'and '3β'. These characteristic changes of respiration are considered to come from an increase in respiration irregularity which is caused by a struggle against drowsiness.

%Blink-A was highest at '1α' and significantly decreased at other states, which is reasonable from the definition (alert blinks). The effect of 'effort' is observed in an increase of *%Blink-A* and *%Blink-B* at '2β' and '3β', however neither was not statistically significant. *%Blink-C* was lowest at '3α' and *%Blink-D* was significantly higher at '3α' and '3β' than '1α'. *%Blink-E* was significantly higher at '1α' than '2β', '3α' and '3β'. *Blink-N* was lower at '1α' than other states, but not significant.

3.3 Estimation of Arousal States by Physiological Measures

Multiple linear regression analysis was also applied to the data of twelve participants who reached level '3'. Referring the results of ANOVA, indices were selected as The explanatory variables; *EEG-αI, Resp-gf, Resp-diff, %Blink-A, %Blink-C and %BlinkD.*

Table 3. Results of multiple regression analysis

	Group	5					4a					4b	4c
	Index	f27	f42	m23	m45	m46	f33	m22	m37	m44	m51	m34	f52
Arousal Level	EEG-αI	5.85	5.58	0	5.58	0	3.99	3.50	5.32	2.81	3.19	3.71	0
	Rsp-gf	−1.27	−0.56	−1.16	0	0	0.00	−2.52	−1.94	0.00	0.83	−2.31	0
	Rsp-diff	0	0	0	0	0	0	0	0	0	0.13	0	0
	%Blink-A	−0.45	0	−0.06	0	0	−0.16	0	−0.06	−0.08	−0.09	0.16	−0.14
	%Blink-C	−0.24	0	−0.25	−0.37	0	−0.30	−0.17	0	0	0.21	−0.20	0
	%Blink-D	0	0.17	−0.24	0	0.44	0	0	0	0.27	0.15	0	0.37
	intercept	−1.45	−3.22	3.30	−2.98	1.19	−1.94	0.83	−1.56	−1.29	−2.68	1.04	1.29
	R	0.91	0.95	0.70	0.65	0.50	0.72	0.77	0.89	0.71	0.82	0.71	0.80
Effort	EEG-αI	0	0	0	0	1.09	1.64	0	2.26	0	2.19	5.22	1.53
	Rsp-gf	0	0	−0.87	−1.62	0	0	−2.34	−1.78	0	0	0	0
	Rsp-diff	0.15	0.10	0	0	0.06	0.04	0	0	0	0.04	0.27	0
	%Blink-A	−0.12	−0.08	0	0	0	0.07	0	0	−0.05	−0.06	0	−0.11
	%Blink-C	−0.08	−0.13	−0.10	0	−0.02	−0.08	0	−0.07	0	0	0	−0.12
	%Blink-D	−0.09	0	0	0	0	0	0	0	0.12	0.05	−0.40	0
	intercept	0.48	0.35	1.19	2.10	−0.99	−1.45	2.63	−0.04	0.19	−1.94	−4.70	−1.00
	R	0.63	0.73	0.70	0.46	0.51	0.76	0.68	0.76	0.54	0.72	0.73	0.58

Figure 6 shows the examples of the results. The multiple correlation coefficients for ‘arousal level’ were ranged from 0.50 to 0.95, while those for ‘effort’ were ranged from 0.46 to 0.76. Relatively low correlation for ‘effort’ is caused not only by individual differences in physiological changes but also the difficulty in finding ‘effort’ from facial expressions for some participants. Although the number and the combination of selected indices showed great differences between individuals, relatively common indices were selected and the sign of the coefficients were identical in most cases (see Table.3).

4 Conclusion

Two-dimensional model was introduced to define arousal states. Facial images were rated by the combination of ‘arousal level’ and ‘effort’ to wake up. Multi-dimensional physiological indices were studied and characteristic changes depending both ‘arousal level’ and ‘effort’ were observed. Facial scores were estimated using selected physiological as explanatory variables and relatively high multiple correlation coefficients were obtained, which suggests the possibility to use physiological measures as the external criteria. However, the number and combination of selected indices showed great differences between individuals. To obtain common equations is an issue in future.

References

1. Boadle, J.: Vigilance and simulated night droving. Ergonomics 19, 217–225 (1976)
2. O’Hanlon, J.F., et al.: Comparison of performance and physiological changes between drivers who perform well and poorly during prolonged vehicular operation. In: NATO Conference Series III, vol. 3, pp. 87–109 (1977)
3. Keckluind, G., et al.: Sleepiness in long distance truck driving: An ambulatory EEG study of night driving. Ergonomics 36, 1007–1017 (1993)
4. Wierwille, W. W. et al.: Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorisms for detection of driver drowsiness. U.S. Department of Transp., National Highway Traffic Safety Administration. DOT HS 808 247, Final Report (1994)
5. Bittner, R., Smrcka, P., Pavelka, M., Vysoký, P., Pousek, L.: Fatigue indicators of drowsy drivers based on analysis of physiological signals. In: Crespo, J.L., Maojo, V., Martin, F. (eds.) ISMDA 2001. LNCS, vol. 2199, pp. 62–68. Springer, Heidelberg (2001)
6. http://www.awake-eu.org/
7. http://www.aide-eu.org/
8. Ohsuga, M., et al.: The Estimation of Driver’s Arousal State (1) - Based on Facial Expression and Physiological Indices. In: Proc. of the 2008 JSAE Annual Congress, vol. 51(06), pp. 1–4 (2008)
9. Nopsuwanchai, R., et al.: Driver-Independent Assessment of Arousal States from Video Sequences Based on the Classification of Eyeblink Patterns. In: Proc. of the 11th International IEEE Conference on Intelligent Transportation Systems, pp. 12–15 (2008)
10. Ohsuga, M., et al.: Classification of blink waveforms toward the assessment of driver’s arousal levels - an EOG approach and the correlation with physiological measures. In: Harris, D. (ed.) HCII 2007 and EPCE 2007. LNCS (LNAI), vol. 4562, pp. 787–795. Springer, Heidelberg (2007)

The Effects of Visual and Cognitive Distraction on Driver Situation Awareness

Meghan Rogers[1], Yu Zhang[1], David Kaber[1], Yulan Liang[2], and Shruti Gangakhedkar[1]

[1] North Carolina State University, Edward P. Fitts Department of Industrial and Systems Engineering, Raleigh, NC
[2] Liberty Mutual Research Institute for Safety, Hopkinton, MA
{mlroger4,yzhang21,dbkaber,sbgangak}@ncsu.edu,
yulan.liang@libertymutual.com

Abstract. Driver distraction has become a major concern for transportation safety due to the increasing use of in–vehicle devices. To reduce safety risk, it is crucial to understand how fundamental aspects of distracting activities affect driver cognition in terms of roadway situation awareness. This study used a simulator-based experiment to investigate the effects of visual, cognitive and simultaneous distraction on operational and tactical control of vehicles. Twenty drivers participated in the study and drove in following or passing driving scenarios under four distraction conditions (without, with visual, with cognitive, and with simultaneous distraction). Results revealed visual distraction to affect all aspects of driver situation awareness. Cognitive distraction affected comprehension and projection of roadway and vehicle states. Correlation analyses revealed decrements in driver SA due to distraction to be associated with decreases in performance.

Keywords: Driver Distraction, Situation Awareness.

1 Introduction

The popularity of in-vehicle information systems has raised concerns with driver distraction and safety. In a study of crash and near crash events, driver distraction, due to in-vehicle technology, accounted for approximately 25% [1]. To reduce distraction-related crashes and develop countermeasures, there is a need for a clear understanding of how in-vehicle technologies affect driver cognition and behavior. Many studies [1, 2, 3, 4] have demonstrated that distraction degrades driver performance (e.g., lane maintenance, speed control); however, they do not explain how distraction affects cognitive functions, such as situation awareness (SA), and, in turn, vehicle control.

Previous research [5,6] has developed operational definitions of SA in the driving domain and empirically investigated SA responses under various task and distraction conditions. Results have been used as a basis for developing models of the role of SA in transactions between driving task requirements and operator behaviors. In a lead car following task simulation, Ma and Kaber [5] found driver operational behaviors (braking, accelerating) were primarily dependent on perception and to some extent comprehension of own vehicle states relative to traffic. They also found that use of a

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 186–195, 2011.

handheld cell phone (posing both visual and cognitive distraction) degraded driver SA, including perception, comprehension and projection. In a more elaborate driving simulator study, involving operational, tactical (passing) and strategic (navigation) task performance, Jin and Kaber [6] observed tactical tasks to be most dependent on all aspects of driver SA. Strategic tasks required SA but did not load on any one specific aspect. Like Ma and Kaber's [5] results, operational driver behavior was also found to be dependent on perception. In addition, Jin and Kaber [6] found that when roadway hazards were posed to drivers, the importance of SA to performance increased, particularly for operational behavior.

On the basis of this research, the objective of the present study was to identify (or sort-out) the independent and combined effects of visual and cognitive distractions on all aspects of driver SA and to quantify negative effects under normal driving conditions involving operational and tactical behaviors. The research was expected to provide a more complete understanding of a cycle of distraction, SA and performance in driving.

2 Methods

2.1 Participants

Twenty young drivers (10 females and 10 males), between the ages of 16-21 years (*M*=18.8 yrs, *SD*=1.4), were recruited to participate in a driving simulator experiment. We wanted to assess "high-risk potential drivers", who have been targeted by State legislation in terms of driving privilege restrictions, based on expert opinions regarding susceptibility to distraction. All participants were required to have valid driver's licenses, as well as normal vision without wearing glasses or contacts. Participants were compensated for their time at $15/hour.

2.2 Apparatus

We used a high-fidelity STISIM Drive^{TM} M400 simulator that rendered dynamic images of driving scenarios based on driver control input. The simulator integrated three 38-inch HDTV monitors providing drivers with a 135-degree field of view of the roadway (left, center and right of the simulated vehicle). Simulator controls included a modular steering unit with a full-size wheel and adjustable speed-sensitive force-feedback capability, as well as a modular accelerator and brake pedal unit (see Figure 1 for a picture of the simulator setup). The simulator also provided spatialized auditory feedback through a 5.1 surround sound speaker system. The simulator computer systems recorded driver performance data at 20Hz, including steering angle, lane position and vehicle speed, distance to nearby vehicles, etc.

A 12-inch HP tablet computer with integrated display was placed in front and slightly below the right side of the simulator screen to present mock-ups of an in-vehicle navigation system. This was used to present the visual distraction task. The screen was positioned approximately 15 degrees down and 30 degree right of the natural line of sight of participants in using the simulator.

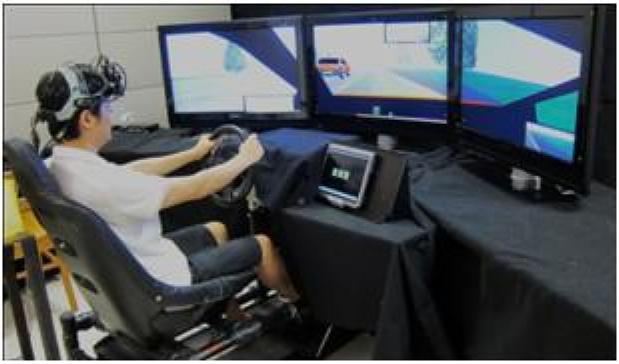


Fig. 1. The STISIM driving simulator and in-vehicle interface

2.3 Experimental Design and Tasks

The experiment was a 2x2x2 within-subject design with two levels of visual distraction (with and without), two levels of cognitive distraction (with and without), and two primary driving tasks. In total there were eight experimental conditions (see Table 1). A lead-car following task required operational control of the simulated vehicle (no maneuvering in traffic or route planning). A passing task required both tactical and operational control, including negotiating complex, multi-lane traffic patterns at high speed. Additional motivation for studying these two task types was a discrepancy in prior research results concerning driver adaptation of behaviors due to cognitive distraction in car following but not in overtaking [7]. Participants drove in eight 8-minute trials with one experimental condition presented in each trial. The sequence of conditions was randomized.

Table 1. Experimental conditions

Primary task	**Cognitive distraction**	**Visual distraction**	
		Yes	No
Passing	Yes	Passing- Simultaneous	Passing- Cognitive alone
	No	Passing-Visual alone	Passing- No distraction
Following	Yes	Following- Simultaneous	Following – Cognitive alone
	No	Following -Visual alone	Following - No distraction

The simulated driving environment was a four-lane interstate highway divided by a green median; two lanes in each direction. The roadway was populated with traffic signs and trees (see Figure 2 for image of simulation display). The speed limit varied between 55mph and 65mph at six virtual locations. In the following task, participants were instructed to follow and maintain a safe distance (using the 4 second rule) to a lead vehicle that drove at the posted limits throughout the trial. The lead vehicle changed lanes 12 times during each trial with a random interval between changes of 20 to 40 seconds; six lane-changes to the right and six lane-changes to the left. The speed limit did not change during the course of lane changes, and there was no interference from traffic.

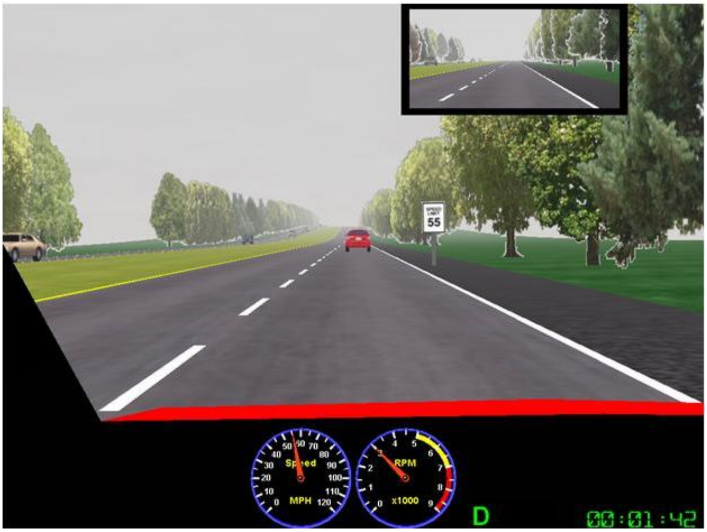


Fig. 2. Simulation Display

In the passing task, participants were instructed to first stay in the right-hand lane and follow a lead vehicle. When the speed of the lead vehicle fell 10 mph below the speed limit, participants were directed to pass, return to the right-hand lane, and follow another lead vehicle. Participants were restricted to passing only one vehicle at a time and instructed to avoid collisions. That is, participants decided whether and when they could make passing maneuvers. They were also instructed to comply with all roadway regulations. A total of six passing events were simulated in each experiment trial with a random interval between passes ranging from 45 to 65 seconds.

There were two types of secondary tasks, including a visual distraction task and a cognitive distraction task. The former simulated use of a navigation aid enroute and diverted driver attention from the road for signal detection while requiring little cognitive effort. Drivers had to identify an upward pointing arrow with a "yellow" background on the tablet display among three arrows pointing in different directions with a default background color of gray. The display was refreshed every 10s. Participants were instructed to respond to each display update and to not make a selection when all arrows had a gray background.

The cognitive distraction task simulated drivers listening to auditory instructions from a navigation system without posing any visual demands and required a verbal response. The instructions described the path of a car traveling on a controlled-access highway loop around a city with exits located in the four cardinal directions (south, east, north, west) and the four intermediate directions (southeast, etc). The car was described as entering the loop from a specific exit and traveling clockwise or counterclockwise. Participants were asked to verbally identify the exit the car would reach (by direction) after passing a certain number of exists. A set of 40 different auditory messages were randomly presented during trials. The cognitive task was delivered every 20 seconds: the message was approximately 5 seconds in duration, and participants were given 15 seconds to respond. During the simultaneous distraction condition, participants performed both the visual and cognitive distraction tasks.

2.4 Procedure

Participants received a 45-minute training session on driving in the simulator, following and passing tasks, and performing secondary tasks while driving. Each participant completed trials under all secondary-task conditions without replication. The secondary task(s) started 30 seconds after the beginning of a trial and ended 15 seconds before the end. Situation awareness was assessed using a real-time probe method. An experimenter sat next to the driver, posing as a vehicle passenger, and asked the subject questions to assess their SA. Questions were posed every 20 seconds for a total of 18 questions per trial. Participants were offered a short break every four trials. The total duration of the experiment was approximately 3.5 hours and evidence of driver fatigue was not observed through statistical analysis of trial order effects.

2.5 Dependent Variables

The dependent variables included driver performance and accuracy in responding to SA queries. Performance measures included speed variability and steering error. Steering error describes how smoothly participants maneuver their vehicle while driving, and was calculated as the absolute difference between the second-order Taylor series expansion prediction of steering angle and the observed angle [8]. A smaller value indicates smoother steering wheel control and less aggressive tactical control. Speed variability determines how variable the driver's speed is to the posted speed limit. We calculated the variance in observed speed during lane changes and passing under each specific limit. Smaller values indicated a better ability to maintain speed close to posted limits. Both measures were recorded when participants performed a lane-change or passed another car. Triggering events for data collection included the lead vehicle steering in the direction of the lane change in the following task or the lead vehicle initiating deceleration in the passing task. Data collection ended when the subject's vehicle changed to the target lane in following or had driven 80 ft. after returning to the right-hand lane in passing. Data was collected on six left-lane changes in following trials and six passes.

Situation Awareness queries were categorized according to the levels of SA defined by Endsley [9] including: perception (Level 1), comprehension (Level 2) and projection (Level 3). Subcategories of queries within levels of SA focused on current roadway objects and background, speed maintenance, traffic patterns, recall of previous roadway objects, roadway topography, time estimates, and horizontal and longitudinal distance estimates. Accuracy of driver responses to queries was analyzed by comparison with ground-truth information on the simulation recorded by experimenters at the time a query was delivered. Aggregate measures of SA for each level were determined on a per subject and trial basis.

2.6 Hypotheses

The present investigation focused on the SA effects of the various types of distraction and driving tasks. Correlation analyses were also conducted on driver SA and performance measures in order to provide insight on how distractions might act through SA to influence operational or tactical driving behaviors. On the basis of previous research by Endsley [9], Matthews et al. [10], and Regan et al. [11] it was hypothesized that low level SA (perception) would reveal a greater influence of visual distractions, and high level SA, requiring complex cognitive processing, would be influenced by both distractions. Based on Jin and Kaber's [6] findings, tactical behavior (the passing task) was expected to require support from higher levels of SA compared to operational driving behavior (the lead car following task). Related to Horrey and Simmons' [7] study, cognitive distractions were expected to have greater influence on the passing task compared to following task, in terms of degraded SA.

3 Results

3.1 Statistical Analysis

Prior to statistical modeling, diagnostics were conducted on the SA and performance data sets using normal probability plots, a test of normality (Shapiro-Wilks test) and tests for constant variance. Subsequently, a multivariate analysis of variance (MANOVAs) was conducted on all response measures for which interrelations were expected, as a means of data reduction. Analyses of variance (ANOVAs) were then conducted on any significant main effects and interaction effects revealed by the MANOVAs across response measures. For further analysis of significant effects, multiple comparison procedures (e.g., Tukey's test) were conducted.

3.2 Situation Awareness

The aggregate responses for all levels of SA met the assumption of the ANOVA; however, some required transformation (arcsine was applied to Level 1 SA and a power transformation was applied to Level 3 SA).

ANOVA results revealed significant effects of primary task type ($F(1,133)=8.78$, $p=0.0036$), visual distraction ($F(1,133)=16.61$, $p<0.001$), a two-way interaction of primary task and cognitive distraction ($F(1,133)=3.96$, $p=0.0487$), a two-way interaction of visual and cognitive distraction ($F(1,133)=8.93$, $p=0.0033$), and a three-way interaction (all main effects; $F(1,133)=24$, $p<0.001$) on Level 1 SA. Mean accuracy of driver perception of roadway states was lower during the passing task (M=0.81, SD=0.16) as compared to following (M=0.87, SD=0.17). A post-hoc analysis of the three-way interaction using Tukey's test revealed the lowest driver

Level 1 SA in passing with cognitive distraction (*M*=0.71, *SD*=0.14), which was significantly different from roadway perception during following with cognitive distraction (*M*=0.93, *SD*=0.14). Passing with cognitive distraction was also significantly worse than driving with visual distraction in either task (following, *M*=0.97, *SD*=0.07; passing, *M*= 0.87, *SD*=0.16), and passing with simultaneous visual and cognitive distraction (*M*=0.86, *SD*=0.15). This latter comparison was surprising as we expected the cognitive demands of the dual-distraction condition to have the greatest implications for driver SA in either task.

Beyond the aggregate measures of accuracy for the general level of SA, it is important to note that certain subcategories of queries revealed SA decrements while others appeared robust to distraction. Regarding the main effect of visual distraction, SA was lower under distraction for queries on traffic pattern (*M*=0.792), recall of previous roadway objects (*M*= 0.825), and horizontal distance estimates (*M*=0.85) as compared to SA on current roadway objects and background (*M*=0.913), speed maintenance (*M*=1) and roadway topography (*M*=0.975).

ANOVA results on Level 2 SA revealed a significant main effect of cognitive distraction (F(1,133)=6.05, p=0.0152) with decrements only occurring for specific types of queries. Accuracy of responses to time and longitudinal distance estimates (*M*=0.35 and *M*=0.3, respectively) were severely degraded compared to no distraction. However, SA on traffic (*M*=0.835), speed maintenance (*M*=0.85), and current roadway objects and background (*M*=0.975) were robust to cognitive distraction. Visual distraction also had a significant effect on Level 2 SA (F(1,133) = 12.06, p=0.0007) with mean accuracy under distraction (*M*= 0.725, *SD*=0.18) being lower than without (*M*=0.813, *SD*=0.16). Results also revealed a significant two-way interaction of primary task type and visual distraction (F(1,133)=15.73, p=0.0001), and a significant two-way interaction of visual and cognitive distractions (F(1,133) = 4.64, p=0.033) on Level 2 SA. In regard to the primary task and visual distraction interaction, during following the distraction did not make a difference; however, driver roadway state comprehension was significantly degraded in passing with visual distraction (*M*=0.68, *SD*=0.17) than during tactical behavior without distraction (*M*=0.87, *SD*=0.15). With respect to the visual and cognitive distraction interaction, vehicle state comprehension relative to the roadway environment was significantly degraded by visual distraction (*M*=0.67, *SD*=0.18) compared to all other conditions [no distraction (*M*=0.81, *SD*=0.15), cognitive distraction (*M*=0.82, *SD*=0.18), and simultaneous distraction (*M*= 0.78, *SD*=0.15)].

Driver accuracy in Level 3 SA (projection of future roadway states) also revealed a significant main effect of cognitive distraction (F(1,133)=16.78, p<0.0001) with worse projection occurring under distraction (*M*=0.72, *SD*=0.21) versus none (*M*=0.83, *SD*=0.15). Visual distraction also had a significant main effect on Level 3 SA accuracy (F(1,133)=6.45, p=0.0123). However, only queries concerning time estimates (*M*=0.6315) and speed maintenance (*M*=0.75) revealed decrements under distraction. SA on traffic (*M*=0.913) and road scenery (*M*=1) were not impacted by the secondary task.

ANOVA results on a Total SA response revealed a significant two-way interaction of primary task type and visual distraction (F(1,133)= 4.3, p=0.0401), and a significant three-way interaction of all three main effects (F(1,133)=5.11, p=0.0254). Post-hoc results on the two-way interaction showed significantly lower total SA for passing with visual distraction (*M*=0.78, *SD*=0.11) versus following (*M*=0.83, *SD*=0.1). Regarding the three-way interaction, it appeared that cognitive distraction during passing was significantly worse for overall SA (*M*=0.77, *SD*=0.10) than visual distraction during following (*M*=0.86, *SD*=0.07).

3.3 Relation of SA and Performance

Correlation analyses were conducted to assess any relation of driver SA with performance, including steering error and speed variability. During lead-car following under visual distraction, results revealed a significant negative correlation between Level 1 SA and steering error (r=-0.5318, p=0.0158) as well as between Total SA and steering error (r=-0.5806, p=0.0073). As SA increased, steering error decreased, suggesting drivers with heightened perception of the roadway demonstrated more stable vehicle control. Similar observations were made on Level 2 SA with steering error (r=-0.2862, p=0.0101) during the passing primary task. Under cognitive distraction during both following and passing tasks, Level 1 SA (perception of roadway states) was negatively correlated with speed variability [(r=-0.458, p=.0488) and (r=-0.4903, p=0.0282), respectively]. This indicated that as SA improved, drivers exhibited more stable vehicle control.

4 Discussion

The two types of driving tasks and two types of distractions significantly interacted in effect on overall driver SA and perception and comprehension of the roadway environment, in specific. The demands of tactical driving behavior in passing led to worse SA than in the following task. Furthermore, the greatest decrements in driver SA due to distraction occurred during the passing task and involved perception and comprehension of the roadway. These findings are in line with Jin and Kaber's [6] results that indicated tactical task performance placed the greatest demands on driver SA.

Visual distraction had a significant effect on all three levels of driver SA while cognitive distraction was only significant for the higher levels of SA (comprehension and projection). These results supported our hypotheses. Most interestingly, we found that for all levels of SA, there were certain driver information requirements that were more sensitive than others to visual and/or cognitive distraction. Highly sensitive aspects of SA in both passing and following tasks included temporal and spatial judgments of roadway events and objects, respectively. The negative effects of visual and cognitive distraction on these SA elements amounted to, on average, a 43% decrease in accuracy in time estimates and a 50% decrease in spatial position estimates. It is possible that certain SA requirements, which we anticipated to be critical to task performance, did not demand cognitive resources to the extent of competition with the visual and/or cognitive distractions.

The relationship between SA and driving performance was in line with expectation. As SA improved, steering error and speed variability decreased, suggesting drivers with heightened awareness of the roadway were more stable in vehicle control. Results on steering error and speed variability served to validate the SA measurement approach, as the general construct of performance was positively associated with SA. This supports the notion of a cycle of driver distraction acting through SA to influence performance. Consequently, the impacts of driver distraction on SA need to be considered in design of new in-vehicle technologies.

5 Conclusion

The findings of this study demonstrate that visual and cognitive distractions impact driver performance by degrading SA. Higher levels of SA are sensitive to cognitive distraction, in particular time estimates and longitudinal distance estimates. Visual distraction appears to affect low level situation awareness, as well as comprehension and projection. This is likely due to perception being a 'building block' for higher level SA, as Endsley [9] described in her theory. In specific, perception of traffic patterns, roadway objects, and horizontal vehicle separation distances were all sensitive to driver visual distraction; whereas, driver projection ability was most sensitive to visual distraction in terms of the timing of roadway events and the need for speed control. Correlation results revealed in-vehicle distractions further impaired drivers' safety by increasing their steering error and speed variability due to a lack of situation awareness.

It should be noted that this study used young drivers (16-21 years old), who have been observed to be more vulnerable to the influence of distraction because of their inexperience and interest in the use of new technologies while driving. Previous research has noted that young drivers represent the highest crash risk population among all drivers [12]. The results of this study provide insight for high-risk drivers, but at the same time may be somewhat exaggerated relative to, for example, the adult experienced driver population. A broader sample population should be investigated in future research. In addition, because no actual threats to safety were posed by driving in the simulated environment, drivers might have engaged in more liberal operational and tactical behaviors, as compared to real-world driving. Beyond this, the secondary tasks used in the study were not directly related to the primary driving task goals or the daily life goals of subjects. Therefore, subjects may not have been motivated to fully engage in the tasks as they would in using an actual cell phone or GPS device.

Acknowledgement. This study was supported by the Edward P. Fitts Department of Industrial & Systems Engineering and College of Engineering at the North Carolina State University. The research was conducted in the Ergonomics Laboratory.

References

1. Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J., Ramsey, D.J.: The impact of driver inattention on near-crash/crask risk: An analysis using the 100-car naturalistic driving study data. Virginia Tech. Transportation Institute (2006)
2. Dingus, T.A., Antin, J.F., Hulse, M.C., Wirewille, W.W.: Attentional demand requirements of an automobile moving-map navigation system. Transportation Research, Part A: General 23A(4), 301–315 (1989)

3. Zhang, H., Smith, M.R.H., Witt, G.J.: Identification of real-time diagnostic measures of visual distraction with an automatic eye-tracking system. Hum. Factors 48(4), 805–821 (2006)
4. Horrey, W.J., Wickens, C.D.: Examining the impact of cell phone conversations on driving using meta-analytic techniques. Hum. Factors 48(1), 196–205 (2006)
5. Ma, R., Kaber, D.B.: Situation awareness and workload in driving while using adaptive cruise control and a cell phone. Int. J. Ind. Ergonom. 35(10), 939–953 (2005)
6. Jin, S., Kaber, D.B.: The role of driver cognitive abilities and distractions in situation awareness and performance under hazard conditions. In: 17th World Congress on Ergonomics, china, August 9-14. Elsevier, Amsterdam (2009)
7. Horrey, W.J., Simons, D.J.: Examining cognitive interference and adaptive safety behaviours in tactical vehicle control. Ergonomics 50(8), 1340–1350 (2007)
8. Nakayama, O., Futami, T., Nakamura, T., Boer, E.R.: Development of a steering entropy method for evaluating driver workload. SAE Technical Paper Series: #1999-01-0892: Presented at the International Congress and Exposition, Detroit, Michigan, March 1-4 (1999)
9. Endsley, M.: Toward a Theory of Situation Awareness in Dynamic Systems. Hum. Factors 37(1), 32–64 (1995)
10. Matthews, G.M., Bryant, D.J., Webb, R., Harbluk, J.L.: Model for situation awareness and driving: Application to analysis and research for intelligent transportation systems. Transportation Research Record 1779, 26–32 (2001)
11. Regan, M.A., Young, K.L., Lee, J.D., Gordon, C.P.: Source of driver's distraction. Driver distraction: Theory, effects and mitigation, pp. 249–279. CRC Press, Taylor & Francis Group, Boca Raton, FL (2009)
12. Stutts, J.C., Reinfurt, D.W., Staplin, L., Rodgman, E.A.: The role of driver distraction in traffic crashes. University of North Carolina, Highway Safety Research Center, Chapel Hill (2001)

Experienced and Novice Driver Situation Awareness at Rail Level Crossings: An Exploratory On-Road Study

Paul M. Salmon[1], Michael G. Lenné[1], Kristie Young[1], and Guy Walker[2]

[1] Human Factors Group, Monash University Accident Research Centre,
Building 70, Clayton Campus, Monash University, Victoria 3800, Australia
[2] School of the Built Environment, Heriot-Watt University, Edinburgh, UK
paul.salmon@monash.edu

Abstract. Poor or degraded situation awareness has previously been identified as a contributory factor in crashes at rail level crossings. Despite this, the concept remains largely unexplored in this context. This paper describes an exploratory on-road study focusing on novice and experienced driver situation awareness whilst negotiating rail level crossings. Participants drove a pre-determined urban route, incorporating two rail level crossings, in an instrumented vehicle. Situation awareness was assessed using propositional networks which were constructed based on content analyses of driver verbal protocols. Differences between drivers' situation awareness were found in terms of the information underpinning it and the integration of this information. It is concluded that, whilst negotiating the two rail level crossings, inexperienced drivers had less efficient situation awareness than experienced drivers. In closing, the implications of this study are discussed along with a series of recommendations for further research in this context.

Keywords: Situation awareness, rail level crossings, on-road studies.

1 Introduction

Crashes at rail level crossings represent a significant road and rail safety problem, both in Australia and worldwide. For example, during 2008, there were 58 collisions between trains and vehicles at rail level crossings in Australia, which led to 33 fatalities and serious injuries [1]. Although fewer in number than general road traffic crashes, incidents at rail level crossings typically engender high levels of trauma, great financial cost, and create great disruption to the road and rail networks.

Despite widespread research focusing on rail level crossing safety [2], rail level crossing system performance remains ambiguous. In particular, elements of driver behavior in this context, such as Situation Awareness (SA), decision making, and driving errors, remain largely unexplored. Further, the causes of rail level crossing crashes remain poorly understood. One factor that has previously been identified as a key causal factor both in general road traffic crashes and rail level crossing incidents, is poor or degraded driver SA. Despite this, SA has not yet been investigated in the context of rail level crossings.

In a driving context, SA can be defined as activated knowledge, regarding road user tasks, at a specific time, within the road transport system. From a road user perspective, this knowledge encompasses the relationships between road user goals

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 196–204, 2011.

and behaviors, vehicles, the road environment and infrastructure. The aim of the exploratory study described in this paper was to investigate driver SA in the level crossing context. Specifically, the authors wished to model driver SA when negotiating rail level crossings with a view to ascertaining what it comprises and comparing it across drivers of differing experience levels. For this purpose, an on-road study in which participants drove a pre-determined route incorporating two rail level crossings was undertaken. This paper presents an overview of the study, including the methodology employed and the findings obtained.

2 Methodology

2.1 Modeling and Analyzing Situation Awareness

Recent studies of SA in complex sociotechnical systems have employed a network analysis based approach, known as propositional networks (Salmon et al, 2009) for modeling and assessing SA. Networks comprising 'information elements' and the links between them are constructed based on content analyses of verbal commentary provided by human operators during task performance. Mathematical analysis of the networks is then used to interrogate their content and structure [3; 4].

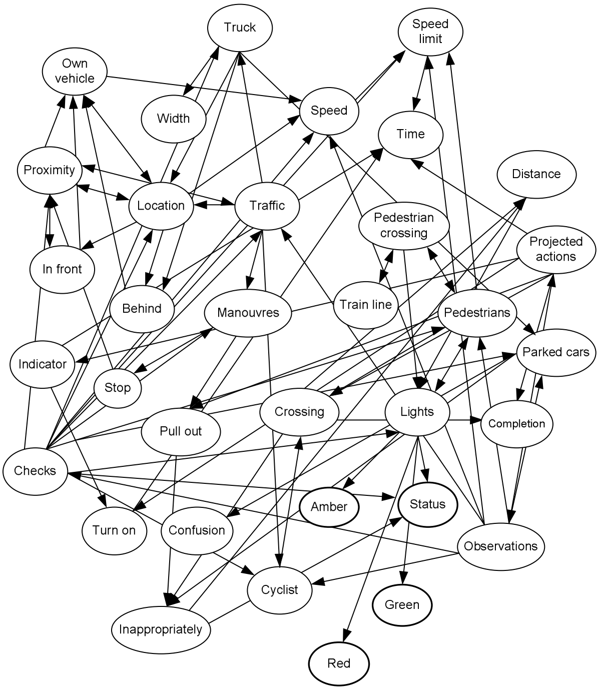


Fig. 1. Propositional network representing participant situation awareness whilst negotiating rail level crossing

In the present study, propositional networks, constructed based on content analyses of verbal protocols provided by participants whilst driving the study route were used to describe participant SA whilst negotiating two rail level crossings. To demonstrate, a propositional network taken from the study is presented in Figure 1; the network presented represents this particular participants SA when negotiating one of the level crossings.

2.2 Design

The study involved an on-road study whereby participants drove an instrumented vehicle around a pre-defined urban route which included two rail level crossings. Two observers sat in the vehicle, and participants were required to provide a concurrent verbal protocol as they drove. Post drive, propositional networks, constructed based on content analysis of participant verbal protocols, were used to depict participant SA at each level crossing. Each network was then analyzed using network statistics.

2.3 Participants

Twenty-three drivers (15 males, 8 females) aged 19-59 years (mean = 27.9, SD = 11.1) took part in the study. Participants were allocated into an experienced driver group or a novice driver group based on driving experience. The experienced driver group comprised fourteen participants (mean age 32.7 years) who held a valid Full license and had an average of 14.3 years solo driving experience (SD = 11.8). The novice driver group comprised nine participants (mean age 19.8 years) who held a valid probationary license and had an average of 1.8 years solo driving experience (SD = 0.8).

2.4 Materials

A demographic questionnaire was completed using pen and paper prior to the drive. A 21km urban route incorporating a short practice route and four active rail level crossings (two of which are focused on in the present paper), both with boom gates, bells and flashing light controls, was used for the on-road study (see Fig 2). Participants drove the route using MUARC's On-Road Test Vehicle (ORTeV), which is equipped to collect three types of data: vehicle-related data, driver-related physiological data, and eye tracking data. A Dictaphone was used to record participant verbal transcripts during the drive. In-vehicle observers used a driver behavior pro-forma to record events and behaviors of interest during the drive.

2.5 Procedure

Following a short briefing session, participants took part in a short VPA training session in which they were given an overview of the approach and opportunity to practice providing concurrent verbal protocols. Following this, participants were taken to the ORTeV and told to set themselves in a comfortable driving position. Two observers were present in the vehicle throughout the drive. Upon commencing the drive, participants first completed a practice route whilst providing verbal protocols. Upon reaching the end of the practice route, participants were informed that the test had begun and that data collection had now commenced. At this point the Dictaphone was switched to record. On-route, the observer located in the front passenger seat provided directions. Drivers provided verbal protocols constantly throughout the drive.

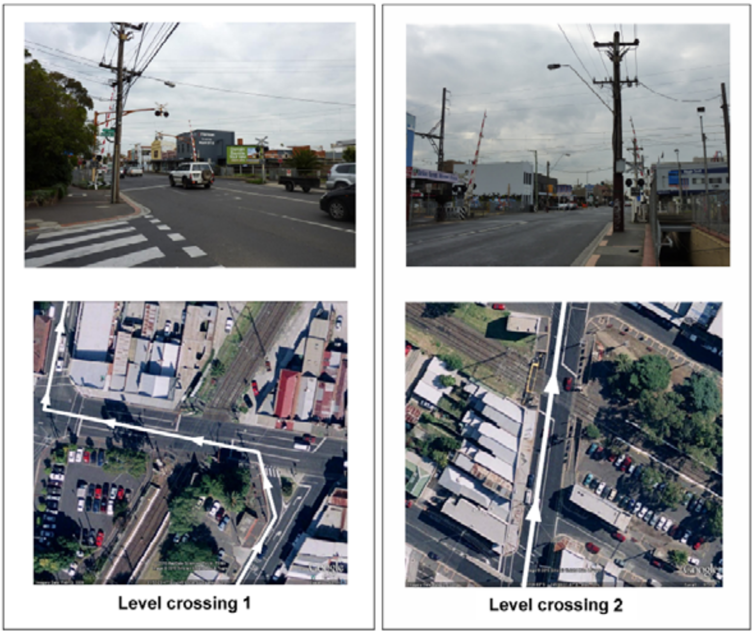


Fig. 2. Bird's eye and approach views of rail level crossings involved in study

Participant verbal protocols were transcribed using Microsoft Word. For data reduction purposes, extracts of each participant's verbal transcript covering the approach to and negotiation of both level crossings were taken from the overall transcript. The extracts were taken based on set points located in the road environment prior to and after each level crossing. One analyst with significant experience in the construction of propositional networks then performed a content analysis on each transcript in order extract information elements and the links between them for each participant at each level crossing. This information was then used to construct two propositional networks for each participant representing their SA at each level crossing. A software package, Agna, was then used to analyze the networks mathematically.

3 Results

Out of the 46 rail level crossing encounters (two for each participant) only 5 involved either rail level crossing in an activated state (i.e. barriers down and lights flashing with a train approaching). Since this proportion is too low for a meaningful comparison of driver SA at activated versus not activated rail level crossings, data for the activated rail level crossings was removed from the analysis. The analysis presented therefore includes the data for the 41 non-activated rail level crossing encounters.

3.1 Network Analysis

A total of 46 networks (1 for each rail level crossing per participant) were produced from the verbal transcripts. The structure and content of propositional networks was analyzed using various network statistics. A brief description of each metric is given below, followed by the results derived from each form of analysis.

Network density represents the level of interconnectivity of the network in terms of links between information elements and is expressed as a value between 0 and 1, with 0 representing a network with no connections between information elements, and 1 representing a network in which every information element is connected to every other information element [4]. In the context of SA, higher levels of interconnectivity indicate an enhanced, richer level of SA with more well integrated information elements and more linkages between them. Poorer SA would be embodied by a lower level of interconnectivity between information elements, since the information underpinning SA is not well linked or integrated.

Diameter is used to analyze the connections between concepts within networks and also the paths between the concepts [4]. Greater diameter values are indicative of more concepts per pathway through the network [4]. Denser networks have smaller values since the routes through the network are shorter and more direct. For SA, lower diameter scores are indicative of better SA, since the holder is able to generate awareness through the linkage of information elements, whereas higher diameter scores are indicative of a model of the situation comprising more information elements but with less links present between them.

Finally, the content of the networks, that is what SA comprises in terms of specific information elements, is analyzed using the *sociometric status* metric. This provides a measure of how ‘busy’ a node is relative to the total number of nodes present within the network under analysis [5]. In relation to SA, the sociometric status of each information elements represents how prominent they are in terms of connections with other information elements within the network. Nodes with high values are highly connected to other nodes in the network, whereas nodes with low values have low connectedness with other nodes in the network. Accordingly the sociometric status has been previously used as a way of identifying the key information elements underpinning SA in complex sociotechnical systems [3].

Network Density and Diameter. The mean density and diameter values for the participants’ networks at both level crossings are presented in Figure 3.

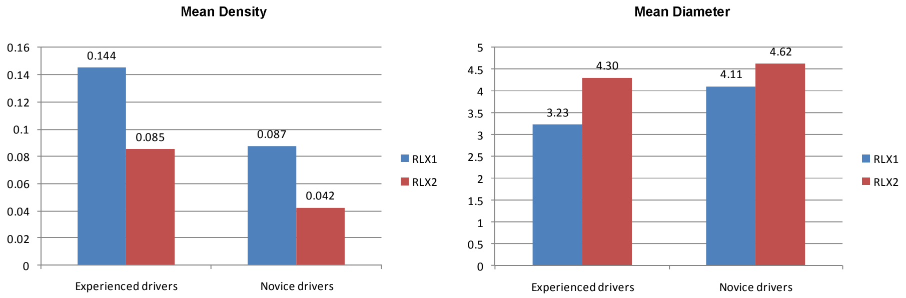


Fig. 3. Mean density and diameter scores from full and probationary license holder networks

At the first rail level crossing, the SA networks derived from the experienced driver group had a mean density of 0.144, whereas the networks derived from the novice driver group had a mean density of 0.087. At the second rail level crossing, the mean density scores were 0.085 for the experienced drivers and 0.042 for the inexperienced drivers. Experienced driver SA, when characterized by propositional networks, was more interconnected at both rail level crossings than that of inexperienced drivers.

The inexperienced driver groups had greater mean diameter scores at both rail level crossings. At the first crossing, the experienced driver networks had a mean diameter score of 3.23 compared to 4.30 for the novice drivers. At the second crossing the experienced drivers had a mean diameter score of 4.62 compared to 4.11 from the novice drivers. These results indicate that the novice drivers' SA networks comprised more information elements but with fewer connections between them compared to experienced drivers.

Sociometric Status. The sociometric status metric was used to identify the key information elements underpinning each participant's SA. Key information elements are defined as those that have salience during task performance, salience being defined as those information elements that act as hubs to other information elements. For each network, those information elements with a sociometric status value above the mean sociometric status value for that network were defined as key information elements [3]. Generic categories of information elements were then devised and a frequency count of the key information elements in each category group was undertaken. The key information elements for both driver groups at rail level crossing 1 are presented in Figure 4.

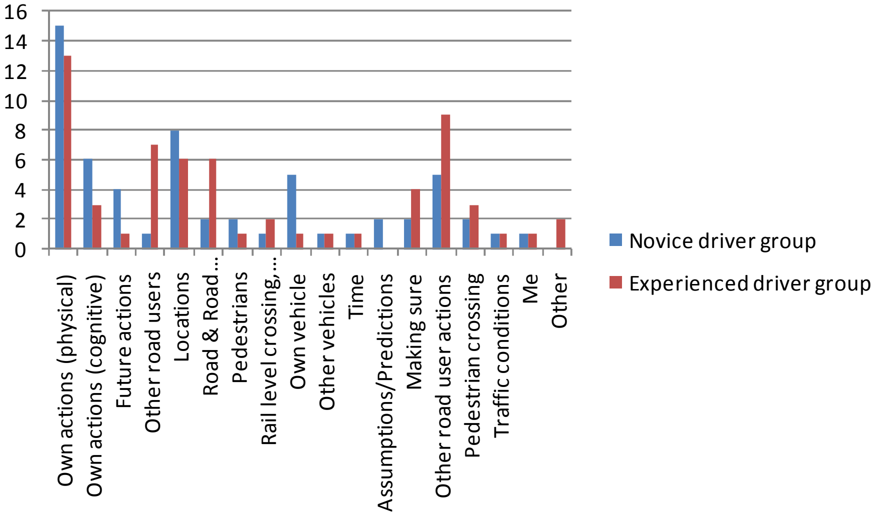


Fig. 4. The frequency of key information elements for novice and expert drivers at rail level crossing 1

Whilst driving through rail level crossing 1, the most common key information elements from both groups were elements related to the participants own actions, both physical (i.e. vehicle control actions) and cognitive (i.e. observing, checking, thinking). The novice driver group had twice as many key information elements relating to drivers own cognitive actions (i.e. observing, checking, thinking) than the experienced driver group did, and had more future actions (i.e. forecasted driving actions), own vehicle and assumptions/predictions (i.e. predictions regarding other road user behaviors) related key information elements than the experienced driver group did. The experienced drivers had more key information elements relating to other road users (e.g. other drivers, cyclists), other road user actions, the road and road infrastructure (e.g. traffic lights, road markings).

The key information elements for both driver groups at rail level crossing 2 are presented in Figure 5.

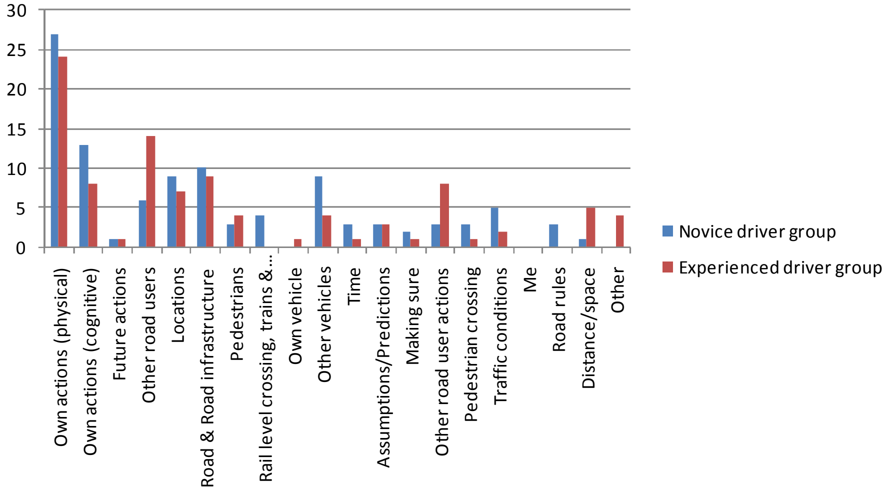


Fig. 5. Key information elements for novice and expert drivers at rail level crossing 2

A similar pattern was found at rail level crossing 2, with the most common key information elements from both groups being elements related to the participants own physical actions. Again, the novice driver group had more key information elements relating to their own cognitive actions, locations, the road and road infrastructure, rail level crossing and infrastructure, other vehicles, time, traffic conditions, and road rules than the experienced driver group did. The experienced driver group had more key information elements relating to other road users, other road user actions and space in the road environment.

4 Discussion

This exploratory study aimed to investigate experienced and novice driver SA whilst negotiating rail level crossings. The findings indicate that, when driver SA is described using propositional networks, experienced driver SA was more interconnected in terms of linkages between information elements extracted from the road environment, and comprised less information elements that did novice drivers' SA networks.

The content of SA, embodied by the key information elements underpinning drivers' SA networks, was also found to be different across the two driver groups. Experienced drivers' key information elements were related to other road users and their actions, and the road and road infrastructure, whereas the novice drivers' key information elements were related more to their own current and future physical and cognitive driving actions, their own vehicle, and predictions regarding the driving environment. From this it is concluded that, during the present study at least, experienced driver SA was focused more on other road users and their actions, whereas novice driver SA was more internally oriented, focusing instead on their own driving actions, their own vehicle and location in the driving environment.

The analysis leads us to conclude that novice drivers experience rail level crossings differently to more experienced drivers. Extracting more raw information from the driving situation, novice drivers' situational models were less connected in terms of linkage between different information elements and comprised more distinct information elements. Experienced drivers on the other hand extract less raw information from the driving situation, and are able to generate smaller, more connected situational models. It is concluded therefore that novice driver SA was less efficient than the experienced drivers' SA during the study. Less (information elements) means more (SA) with respect to experienced driver SA in the rail level crossing context.

The findings presented have significant implications for driver training. Advanced driver training programs have previously been found to enhance driver SA in terms of increased connectivity between information elements [e.g. 6; 7]. Further, poorly developed schema has been identified as one potential cause of diminished SA within complex sociotechnical systems [3]. Since the novice driver group were found to have lower levels of interconnectivity when negotiating rail level crossings, increased exposure to rail level crossings during driver training programs is advocated by these authors as a way of developing novice driver schema related to rail level crossings.

As an exploratory study this research did have some key limitations. First, the lack of active rail level crossings encountered by participants was problematic; however, this was a natural risk associated with conducting the study semi-naturalistically on road and the low number of active crossings meant that the findings were not influenced adversely. Second, the use of one analyst to build the propositional networks could be criticized on the grounds of concerns over the reliability of the propositional network methodology; however, the analyst in question has significant experience in constructing such networks in a range of different domains.

Acknowledgement. This research was funded by the Commonwealth of Australia, through the Cooperative Research Centre for Advanced Automotive Technology. Dr Paul Salmon's contribution to this research and article was funded via an Australian National Health and Medical Research Council (NHMRC) Post doctoral Public Health training fellowship.

References

1. Australian Transport Safety Bureau. Australian Rail Safety Occurrence Data to 31st Report number RR-2008-011(2) (2009)
2. Edquist, J., Stephan, K.L., Wigglesworth, E., Lenné, M.G.: A literature review of human factors and safety issues at Australian level crossings. Monash University Accident Research Centre report (2009)
3. Salmon, P.M., Stanton, N.A., Walker, G.H., Jenkins, D.P.: Distributed situation awareness: advances in theory, measurement and application to teamwork. Ashgate, Aldershot (2009)
4. Walker, G. H., Stanton, N. A., Salmon, P. M. Cognitive compatibility of motorcyclists and car drivers. Accident Analysis & Prevention (December 9, 2010) (in Press)
5. Houghton, R.J., Baber, C., McMaster, R., Stanton, N.A., Salmon, P.M., Stewart, R., Walker, G.H.: Command and control in emergency services operations: a social network analysis. Ergonomics 49, 1204–1225 (2006)
6. Stanton, N.A., Walker, G.H., Young, M.S., Kazi, T., Salmon, P.M.: Changing drivers minds: the evaluation of an advanced driver coaching system. Ergonomics 50(8), 1209–1234 (2007)
7. Walker, G.H., Stanton, N.A., Kazi, T.A., Salmon, P.M., Jenkins, D.P.: Does advanced driver training improve situation awareness? Applied Ergonomics 40(4), 678–687 (2009)

Influence of Brightness and Traffic Flow on Driver's Eye-Fixation-Related Potentials

Yoshihisa Terada[1], Koji Morikawa[2], Yuji Kawanishi[3],
YongWook Jeon[4], and Tatsuru Daimon[5]

[1] Tokyo R&D Center, Panasonic Corporation,
600 Saedo-cho, Tsuzuki-ku, Yokohama, Japan
terada.yoshihisa@jp.panasonic.com
[2] Advanced Technology Research Laboratories, Panasonic Corporation,
3-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
morikawa.koji@jp.panasonic.com
[3] Graduate School of Science and Technology, Keio University,
3-14-1 Hiyoshi, Kouhoku-ku, Yokohama-shi, Kanagawa, Japan
kikou.basketbaii.9@z8.keio.jp
[4] Division of Industrial and Information Systems Engineering, Ajou University,
KoreaSan 5, Wonchen-dong, Yeontong-gu, Suwon, Korea
jyw0673@gmail.com
[5] Faculty of Science and Technology, Keio University,
3-14-1 Hiyoshi, Kouhoku-ku, Yokohama-shi, Kanagawa, Japan
daimon@ae.keio.ac.jp

Abstract. This paper investigates the influence of environmental factors and driver distraction on the eye-fixation-related potential (EFRP) of drivers. Brightness and traffic conditions were set up as environmental factors in experiments using a motion-based driving simulator, and several cognitive tasks were given simultaneously to the participants while they simulated driving. The results of this experiment show that brightness and traffic flow do not affect the EFRP. This shows that EFRP is a stable index of driver distraction.

Keywords: Electroencephalogram, Eye-fixation-related potentials, Distraction.

1 Introduction

In Japan, over 770,000 traffic accidents are reported per year [1]. About 31% of all traffic accidents are rear-end collisions. Driver-related factors in rear-end collisions include drivers taking their eyes off the road, drowsiness, inebriation, sudden illness, improper vision day dreaming, and speeding [2]. To decrease the number of car accidents, it is important to develop a safe-driving support system that can judge whether the driver is in a suitable state.

In this paper, we focus on driver distraction, which is a major factor in rear-end collisions. A distracted state is defined as a situation in which the driver's attention is focused on non-driving tasks such as conversations or thinking, and the driver is not paying sufficient attention for safe driving.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 205–213, 2011.

Eye-fixation-related potentials (EFRPs) have been proposed as an index for levels of visual attention. An EFRP is a kind of electroencephalogram (EEG) potential that is elicited after the end of a saccadic eye movement. When an eye fixation begins, positive amplitude is observed for around 100 ms around the back of the head. This positive amplitude is called a lambda response, or P100. The amplitude of a lambda response increases according to the level of attention paid to an object [3], [4].

Some studies have reported the assessment of driver distraction levels based on EFRP. Nakada et al. [5] reported the influence of driver distraction and road situations (intersections and straight roads) on EFRP. During a complex driving situation, there is a difference in peak amplitude of the lambda response between the driver's attention levels. Yagi et al. [6] reported that illuminations increase the amplitude and delay the latency of the lambda response in dark conditions. However, the influence of road brightness, such as during day and nighttime, as well as traffic flow, is not clear.

The purpose of this study is to clarify the influence of brightness and traffic flow on EFRP. A series of experiments using a driving simulator (DS) were conducted, and the results showed the stability of EFRP as an index of distraction.

2 Methods

2.1 Main Task and Subtask

A dual-task paradigm is used for examining states of distraction while driving. A main task and subtask are given to the participants concurrently. The main task is driving, and the subtask is a cognitive activity.

Driving Task. An urban driving course was prepared, and the driver was allowed to accelerate freely within the speed limit. The driving course was made to simulate an existing city (Yokohama, Japan). The conditions of the course (congestion, number of turns, etc.) were set equally for each condition. Participants had to drive along the route using a car navigation system. Since the route was displayed only on the screen of the car navigation system without audio guidance, the participant had to check the screen periodically.

Cognitive Task. To distract the participants, question-and-answer tasks were given to the participants as cognitive tasks. The questions were designed for two difficulty levels. The more difficult questions were regarding declarative knowledge of geography, such as names of the prefectures in Japan, or names of countries beginning with the letter "A." This condition was set as a distracted state and required significant attention for thinking and memory recall. The easier questions were regarding personal information, such as the driver's name and age. This condition was set as a concentrate state as the participants could answer such questions without being distracted from their driving.

The recorded questions were presented by the experimenter. The participants were required to answer orally as soon as possible. Once the participant answered, or the experimenter judged that the question was too difficult to answer, the next question was presented.

2.2 Environmental Factors

Brightness. Two conditions, bright and dark, were prepared as the brightness factors. Luminance at the driver’s eye position was measured using a photocell light meter.

Luminance under a bright condition was 40.0 lx, and 2.2 lx under a dark condition. Figures 1 and 2 show examples of the driver’s view point.

Traffic Flow. Crowded and sparse conditions were prepared as traffic flow factors. Under a crowded condition, about eight moving objects, including cars, pedestrians and bicycles, were placed for 50 meters. Under sparse conditions, there were only two or fewer moving objects per 50 meters. Figures 3 and 4 show examples of the driver’s view point.

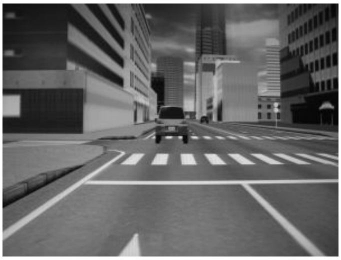


Fig. 1. Bright condition

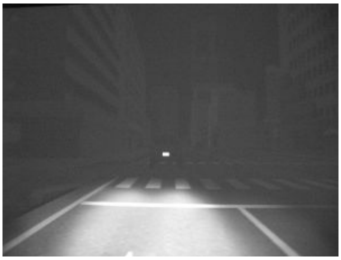


Fig. 2. Dark condition

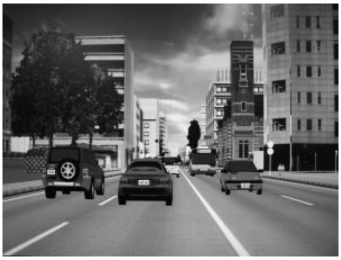


Fig. 3. Crowded condition

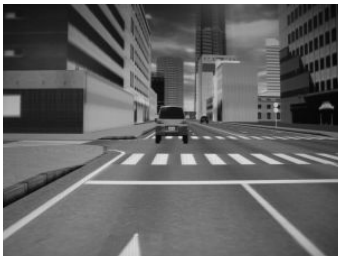


Fig. 4. Sparse condition

2.3 Experimental Conditions

Table 1 shows the different experimental conditions. Six experimental conditions were set up based on a combination of cognitive task, brightness, and traffic flow conditions.

Table 1. Experimental conditions

		Cognitive task	
		Easy questions	Difficult questions
Environmental factors	Bright/ Crowded	Easy/Bright/Crowded	Difficult/Bright/Crowded
	Bright/ Sparse	Easy/Bright/Sparse	Difficult/Bright/Sparse
	Dark/ Sparse	Easy/Dark/Sparse	Difficult/Dark/Sparse

2.4 Measurements

Electrophysiological Recording. Electrophysiological data was recorded using a Polymate AP-216 (TEAC).

Electroencephalograms (EEGs). Electrodes were placed at six scalp locations (Fz, Cz, Pz, Oz, A1, and A2) according to the extended 10-20 system using sintered active electrodes. A2 was used as the initial reference for the recording, and the data were re-referenced to mathematically linked earlobes offline. A ground lead was placed at Fpz.

Electrooculograms (EOGs). A pair of electrodes was placed at the outer canthi of the eyes for a horizontal EOG. Another pair of electrodes was placed at the infra- and supra-orbital of the left eye for a vertical EOG.
Filters. EEGs and EOGs were recorded using a 0.05 Hz high-pass filter (for a constant time of 3 s) and digitized at 500 Hz.

Subjective Measures. Multiple subjective measures such as a simulator sickness questionnaire (SSQ) [7], NASA task load index (NASA-TLX) [8], and a questionnaire regarding the level of driving concentrations were used. The NASA-TLX measures the level of mental workload by using a questionnaire with seven options (min: 1 - max: 9). SSQ measures the level of motion sickness in the driving simulator through the questionnaire. A questionnaire on the level of driving concentration was also taken as a subjective index (min: 1 - max: 7). The participants were asked to fill out a NASA-TLX and concentration questionnaire after each experiment, and an SSQ after only the first and last experiments.

Behavioral Measures. Driving logs, such as speed and position, were recorded from the DS. The standard deviation of distance from the center line was used as an index of driving stability. The value is defined as the distance from the center line to the center of the car, and calculated only on a straightaway.

2.5 Participants

Twenty-two male university students who gave their written informed consent participated in the experiment (average age: 22.2 ± 1.5 years old). Each participant has a driver's license and drives a car on a regular basis; however, this was their first time driving in the DS.

3 Results

Data from sixteen of the participants (average age: 22.4 ± 1.6 years old) were analyzed, and the remaining six participants were excluded from the analysis due to recording errors and motion sickness.

3.1 Behavioral Measures

Figure 5 shows the mean value of the standard deviation in the distance from the center line. The horizontal axis indicates the condition, while the vertical axis indicates the mean value (unit: meter).

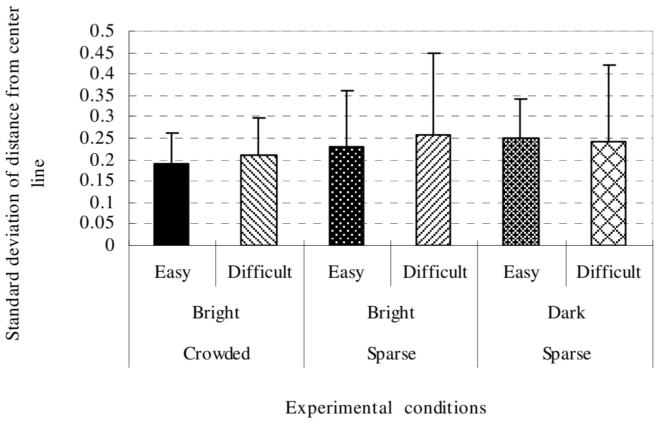


Fig. 5. Standard deviation in the distance from the center line

There was no significant difference among the conditions, as evaluated by Student's *t*-test. Driving behavior seemed to be fairly consistent. This suggests that driving behavior is not affected by cognitive tasks. The driver maintained more than the minimum attention level required to drive under each condition. This confirms that the driver's attention was well controlled during this experiment.

3.2 Subjective Measures

Level of Driving Concentration. Figure 6 shows the mean score of driving concentration. The horizontal axis indicates the conditions, and the vertical axis indicates the score at each level of driving concentration.

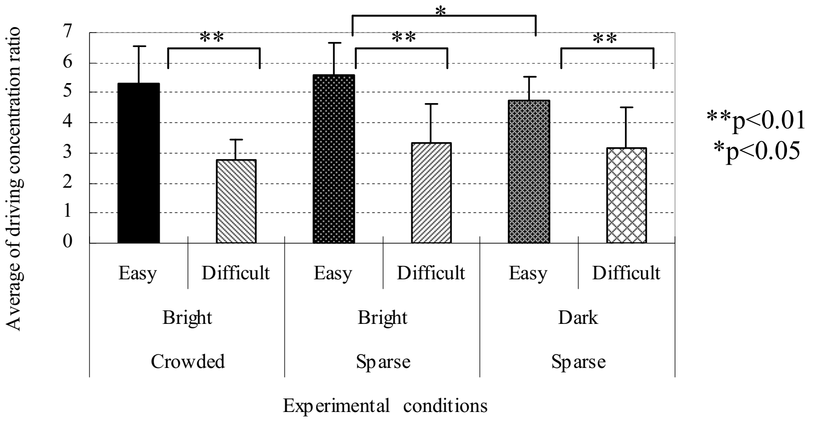


Fig. 6. Subjective scores for driving concentration

The scores were compared and analyzed for three differences using a T test: difficulty, brightness, and traffic flow. For differences in difficulty, the scores were high during easy conditions and low under difficult conditions. The scores were statistically different under easy conditions only. The scores were statistically different during each condition. These results indicate that driving concentration is affected by cognitive tasks, and that many participants cannot concentrate when driving under difficult conditions. For differences in brightness, the scores were higher under bright conditions, but the difference was slight during difficult conditions. The scores were statistically different during easy conditions only. For differences in traffic flow, the scores were low during crowded conditions. However the differences in score were slight between crowded and light traffic conditions. The scores showed no statistical difference.

NASA-TLX. Figure 7 shows a weighted workload (WWL), which is the score weighted by the NASA-TLX. The horizontal axis indicates the condition, while the vertical axis indicates the WWL score.

The WWL scores for the three differences listed above were also compared and analyzed using a T test. For differences in difficulty, the scores for easy conditions were lower than for difficult conditions. The scores were statistically different under each condition. These results indicate that driving concentration is affected by mental workload. For differences in brightness, the scores during dark conditions were a little higher than under bright conditions. The scores were statistically different during easy conditions only. For differences in traffic flow, the scores during crowded conditions were a little higher than during sparse conditions. The scores showed no statistical difference.

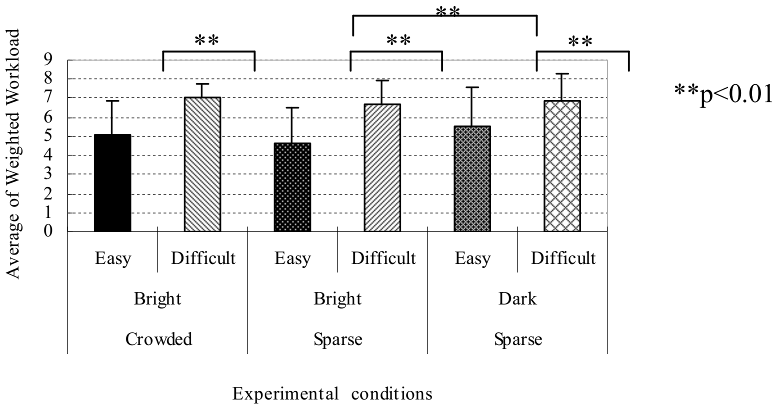


Fig. 7. WWL of NASA-TLX for each experimental condition

The results of these subjective measures show that driver attention is affected not by environmental factors but by driving difficulty.

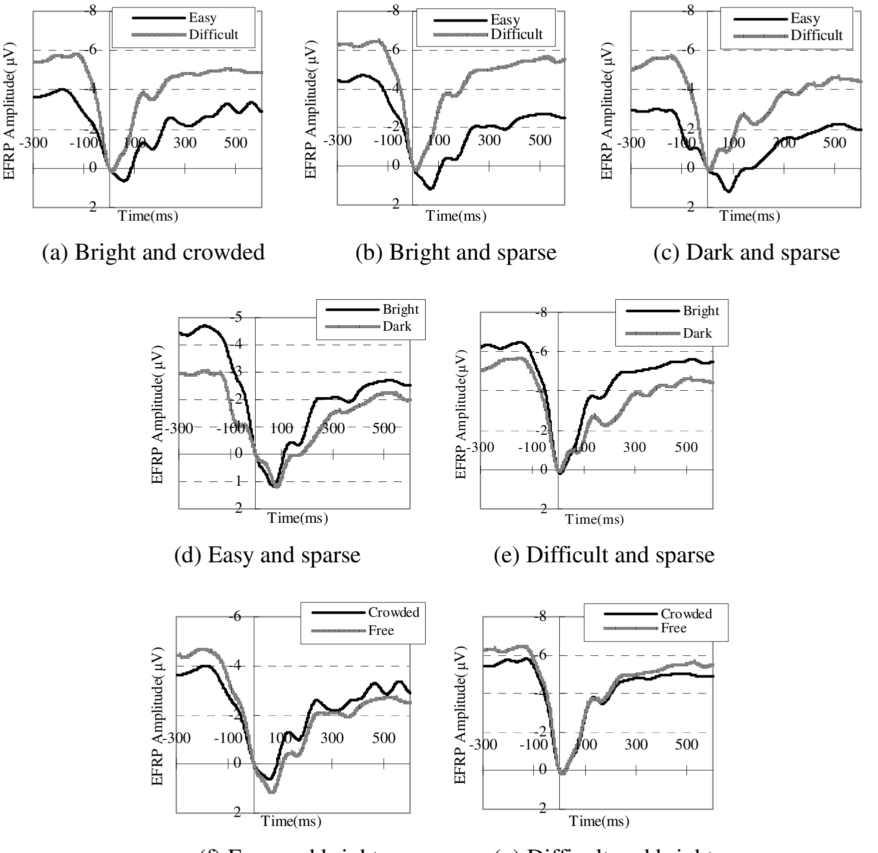
3.3 EFRP

An EOG that continues moving for 20–500 ms at over 3 mV/s is detected as a saccadic eye movement. The EFRPs were obtained by averaging the EEGs whose offsets were determined after the end of saccadic eye movements. The EFRPs were extracted from 300 ms before to 600 ms after the onset of a saccade end.

By evaluating the grand mean wave, we could exclude noises and artifacts (e.g., excessive eye blinking and muscle potentials). The EEG data at Oz were re-referenced mathematically by averaging A1 and A2 offline. A digital 1-15 Hz band-pass filter was also applied.

Because the participants could move their eyes freely during the experiments, the size and frequency of their saccadic eye movements were not equal. The baseline of the EFRP was aligned to the saccade offset amplitude at 0 ms to exclude the effects of eye movement.

Figure 8 shows the grand mean wave of the EFRP for each condition. The horizontal axis indicates the time from the end of a saccadic eye movement, while the vertical axis indicates the amplitude of the lambda response from the obtained EFRP.



(a) Bright and crowded

(b) Bright and sparse

(c) Dark and sparse

(d) Easy and sparse

(e) Difficult and sparse

(f) Easy and bright

(g) Difficult and bright

Fig. 8. Grand mean waveforms of the EFRP under each condition

EFRPs were compared and analyzed using a T test for the three differences, difficulty, brightness, and traffic flow. Differences in difficulty are compared in Figures 8(a), 8(b), and 8(c). Brightness is compared in Figures 8(d) and 8(e), while traffic flow is compared between Figures 8(f) and 8(g). For differences in difficulty, the lambda response increases when the cognitive task is easy, and decreases when the cognitive task is difficult. The amplitude of the lambda response is statistically different under each condition ($p<0.05$, $p<0.01$, $p<0.05$). For differences in brightness, there is no significant difference in the lambda responses between bright and dark conditions. The amplitude of the lambda response shows no statistical difference. For differences in traffic flow, the lambda response under light conditions is larger than during heavy traffic flow during easy conditions. However, during difficult conditions there is no difference in the lambda response. The amplitude of the lambda response also shows no statistical difference.

As a result, the EFRP is affected not by environmental factors in this study but by difficulty only. This result is consistent with the subjective measurement results. This shows that the EFRP has a relation to driver attention. These results suggest that the EFRP is a sensitive index for detecting driver distractions, and is effective even when behavioral differences are not observed.

4 Discussion

Difficulty. For the dual-task paradigm used in this experiment, we requested the participants to engage in driving and cognitive tasks simultaneously. Therefore, the drivers were distracted while conducting difficult subtasks.

The lambda response for easy conditions is larger than for difficult conditions. Nakada et al. [5] reported that the amplitude of the lambda response decreases during high mental workload tasks (2-back cognitive tasks). The results in our study are consistent with their findings. This suggests that cognitive tasks, such as conversations requiring thought and memory recall, cause distractions and reduce driver attention.

Brightness. The amplitude of the lambda response is not statistically different among conditions of brightness. Yagi et al. [6] reported that the amplitude of the lambda response changes based on the brightness conditions, ranging from bright (500 lx, 1300 lx) to dark (5 lx) conditions. The bright and dark conditions in our experiment are 40 lx and 2.2 lx, which were chosen based on the restrictions of the DS. These luminance values are relatively small, and the difference in luminance is also small compared with [6]. Therefore, the level of luminance used in the DS may not be enough to affect the lambda response.

The level of driving concentration and NASA-TLX are affected by brightness during easy conditions only. Under difficult conditions, the driver cannot concentrate on driving when subjected to distractive subtasks. Conversely, under easy conditions, only little attention is taken away by the easy subtask, remaining enough attention for driving. As a result, the driver may react more sensitively to brightness in concentrated state.

Traffic Flow. For all amplitudes of lambda response, questionnaires regarding driving concentration and NASA-TLX showed no statistical differences among traffic conditions. Therefore, traffic flow does not affect driver distractions in this experiment.

The crowded conditions used in our experiment contained eight moving objects per 50 meters, which are usual conditions on urban roads. Additional experiments under more difficult traffic situations, such as unexpected vehicle movement or rushing pedestrians, are necessary to discuss the effects of traffic flow on EFRPs.

5 Conclusion

In this study, we examined how EFRPs are influenced by distractions, brightness levels, and traffic flow. We confirmed that the lambda response decreases when drivers lose concentration caused by thought and memory recall. We also showed that variations in luminance and traffic flow did not affect the EFRPs in this experiment. This shows the stability of the EFRP as an index for driver distraction, and how an EFRP may be used without considering environmental changes such as brightness and traffic levels.

As future work, we need to examine the effects of distractive environmental changes in further detail. Experiments under sunlight or harder traffic conditions should be conducted. We are also aiming toward a feasibility study by extending the attributes that may cause the driver distraction. For example, the separation of arousal and distraction is important for safe driving support.

References

1. Institute for Traffic Accident Research and Data Analysis. In: 2008 Annual Report about Statistics of Traffic Accident, pp. 1–3, Tokyo (2009) (in Japanese)
2. Ohori, T.: Why Does the Rear-end Collision Accident Occur? Institute for Traffic Accident Rsearch and Data Analysis Information 43, 1–8 (2003)
3. Yagi, A.: Visual Signal Detection and Lambda Responses. Electroencephalography and Clinical Neurophysiology 52, 604–610 (1981)
4. Daimoto, K., Suzuki, M., Yagi, A.: Effects of a Monotonous Tracking Task on Eye Fixation Related Potentials. The Japanese Journal of Ergonomics 34(1), 59–65 (1998)
5. Nakada, T., Terada, Y., Morikawa, K., Jeon, Y., Daimon, T.: A Study about Influence on Eye-fixation Related Potential by Driver’s Distraction and Driving Situation. Transactions of Society of Automotive Engineers of Japan (65-10) (2010)
6. Yagi, A., Ito, M., Hirao, N.: Eye Fixation Potentials at Changed Illuminations in the Working Space. In: Annual Conference of the Illuminating Engineering Institute of Japan, vol. 29, p. 260 (1996)
7. Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.L.: Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. International Journal of Aviation Psychology 3(3), 203–220 (1993)
8. Miyake, S., Kumashiro, M.: Subjective Mental Workload Assessment Technique-An Introduction to NASA-TLX and SWAT and a Proposal of Simple Scoring Methods. Japan Ergonomics Society 29(6), 399–408 (1993)

Cognitive Compatibility of Motorcyclists and Drivers

Guy H. Walker[1], Neville A. Stanton[2], and Paul M. Salmon[3]

[1] School of the Built Environment, Heriot-Watt University, Edinburgh, UK
[2] Transportation Research Group, University of Southampton, Southampton, UK
[3] Accident Research Centre, Monash University, Victoria 3800, Australia
G.H.Walker@hw.ac.uk

Abstract. Incompatibility between different types of road user is a problem that previous research has shown to be resistant to a range of interventions. Cars and motorcycles are particularly prone to this. Insight is provided in this paper by a naturalistic method using concurrent verbal protocols and an automatic, highly reliable semantic network creation tool. Analysis of the structure and content of the semantic networks reveals a greater degree of cognitive compatibility on faster roads such as motorways, but evidence of more critical incompatibilities on country roads and junctions. The results are discussed in terms of practical measures such as road signs which warn of events behind as well as in front, cross-mode training and the concept of route driveability.

1 Introduction

It makes intuitive sense that motorcyclists will interpret the same road situation differently to car drivers. The question this paper explores is whether this can really be regarded as the case, and if so, whether car drivers and motorcyclists can be regarded as cognitively compatible? If they are compatible then safety interventions concerned with the objective state of the situation (i.e. increased rider conspicuity) will be more likely to have an effect. This is because drivers will be interpreting the situation in a way that is already favourable to the anticipation of other road users. If they are incompatible then a more nuanced approach might be needed. In this case, no matter how visible a rider may be, if the driver is operating in a situation which is generating a strong stereotypical response unfavourable to the observation of other road users, then in order to improve safety the mental representation of the situation becomes as important as its objective state. We refer to this as cognitive incompatibility.

A generic example of cognitive incompatibility might be described by Norman (1990) as a 'gulf of evaluation'. This describes a person's attempts to make sense of their context and how it matches their expectations and intentions. In Norman's examples, designers and users of a system bring to bear different cognitive models of a system based on their own understanding of it, leading to incompatibilities between what the designer expects and what the user wants. Replace 'designers and users' with 'motorcyclists and car drivers' and it is apparent that similar 'gulfs of execution' can exist in terms of how identical road situations are interpreted, and what those situations might 'afford' for different road users. The concept of 'affordances' reflects the Gibsonian (1979) idea that a relationship exists between people and their

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 214–222, 2011.

immediate context and the Neisserian (1976) concept that the environment is sampled, which in turn modifies behavior, which in turn guides further exploration. Affordances infer that the perceived state of a given context is as important as its objective state. Exploration suggests that perceived states are dynamic and evolving. In this paper semantic networks are used as a way of representing such states.

Semantic networks are based on the long held belief that all knowledge is in the form of associations and represent concepts by depicting linked nodes in a network (Eysenck & Keane, 1990). Within a semantic network each node represents an object. Nodes are linked with edges typically specified by verbs or by analyzing the closeness of concepts using some form of Thesaurus learning algorithm. A variation on this theme is 'concept maps' (Crandall. Klein & Hoffman, 2006). Concept maps are based on Ausubel's theory of learning (Ausubel, 1963) which suggests that meaningful learning occurs via the assimilation of new concepts into existing concepts within the mind of the learner. With close similarities in both approaches, Anderson (1983) proposed 'propositional networks' to describe activation in memory. Salmon (2009) and Stanton (2009) have since extended this approach into the realm of situational awareness and have anchored it successfully to a generative model of cognition (e.g. Neisser's perceptual cycle, 1976) and to Schema Theory (Bartlett, 1932). Schema theory describes how individuals possess mental templates of past experiences which are mapped with information in the world to produce appropriate behavior. A schema is rather like a mental template, which is neither completely new behavior nor merely a repetition of old behavior, but is behavior which is generated from a familiar set of initial conditions, both mental and physical. Schema theory offers an explanation for the paradoxical case described above in which more experienced drivers seem to have greater degrees of cognitive incompatibility with motorcyclists. In this case, because cars are more numerous than motorcycles (in the UK at least), repeated experience with the latter may contribute towards mental templates which generate strong stereotypical behaviours potentially unfavourable to the latter.

This question will be explored in the current article by creating semantic networks based on verbal commentaries provided by car drivers and motorcyclists using a highly reliable automated process called Leximancer™. Potential incompatibilities will be revealed by differences in the structure and content of these networks.

2 Method

2.1 Participants

Twelve participants took part in the study using their own vehicles. They were comprised of six car drivers and six motorcyclists. All participants held a valid UK driving licence with no major endorsements, and reported that they drove approximately average mileages per year for their vehicle type. The participants fell within the age range of 20 to 35 years old. Mean driving experience was 5.83 years for the car drivers and 5.33 years for the motorcyclists.

2.2 Design

The experiment is exploratory and based upon a naturalistic on-road driving paradigm where individuals use their own vehicles around a defined course on public roads. The experimenter travelled in the front passenger seat during the observed runs in the cars, or followed on another motorcycle during the observed runs with the motorcyclists. This controlled for the possible effects of observation upon driving behaviour. Drivers/riders were required to provide a concurrent verbal protocol as they traversed the road course, which was then analysed using a text analysis tool called Leximancer (see Smith, 2003). This enabled differences in textual and thematic content to be systematically analysed, and the structure of the verbal protocol to be mapped using semantic networks. These output are dependant upon two independent variables: vehicle type and road type. Vehicle type has two levels: car or motorcycle. Road type has six levels: motorway(freeway), major road, country road, urban road, junction and residential road. Controlling measures were self-report questionnaires of driving style, recordings of average speed and time, and demographic data. All experimental trials took place at defined times to control for traffic density and weather conditions.

2.3 Materials

Six cars (a Volkswagen Golf TDi, Audi TT, Toyota Tercel, BMW 325i, Volkswagen Golf CL and Peugeot 309 GLD) and six motorcycles (a Triumph Daytona 900, Suzuki TL1000R, BMW R1100GS, Laverda 750 Formula S, Suzuki GSX400F and Honda CBX750) were used in the study. Car drivers were audio recorded whilst they drove using a microphone and laptop computer. Motorcyclists were audio recorded using a microphone mounted in the chin-piece of their crash helmet and a digital recording device carried on their person. An identical set up was used for the accompanying rider.

The on-road route is contained within the West London area of Surrey and Berkshire and was 14.5 miles in length not including an initial three mile stretch used to warm up participants. The route is comprised of one motorway section (70 mph speed limit for 2 miles), seven stretches of major road (50/60 mph speed limits for 6 miles), two stretches of country road (60 mph speed limit for 3 miles), three stretches of urban roads (40 mph limit for 2 miles), one residential section (30 mph limit for 0.5 miles), and fifteen junctions (>30 mph speeds for 1 mile). Experimental runs took place at 10:30 in the morning and 2:30 in the afternoon (Monday-Thursday) and 10:30 on Friday. These times avoided peak traffic hours for the area, and all runs were completed in dry weather.

2.4 Procedure

Formal ethical consent was obtained from all participants before the study commenced with particular emphasis on control of the vehicle and safety of other road users remaining the participants' responsibility at all times. An instruction sheet on how to perform a concurrent verbal protocol was read by the participant, and the experimenter provided examples of the desired form and content. In the case of the motorcyclists, they were further instructed that the experimenter would follow them on another motorcycle in an offset road position. They were instructed to use their mirrors as normal and watch for directional indications from the experimenter and to act upon them.

There then followed a warm-up phase. A three mile approach to the start of the test route enabled the participants to be practised and advised on how to perform a suitable concurrent verbal protocol. This involved providing suggestions and guidance from the passenger seat, or in the case of motorcyclists, pulling over to review the audio transcript and advise where necessary. All participants were able to readily engage in this activity and minimal advice was needed.

During the data collection phase the experimenter remained silent aside from offering route guidance and monitoring the audio capture process. For the motorcyclists the experimenter followed at a safe distance, remaining in the lead rider's rear view mirrors by riding in an offset position, and using their own indicators to guide the participant around the route. Small signs were placed at the roadside to serve as boundaries between road types. These signs were captured on video during observed runs with car drivers, which in turn enabled the verbal transcript to be suitably partitioned. In the case of motorcyclists, the observer carried an audio capture device synchronized with that carried by the participant. When the participant was observed to pass a roadside marker the observer noted this verbally. The two transcripts were combined to allow the data to be partitioned as before.

The verbal protocol data was then treated with Leximancer™, a software product that automates the process of semantic network creation. Six main stages are performed in order to transform verbal transcripts into semantic networks:

1. Conversion of raw text data (definition of sentence and paragraph boundaries etc.).
2. Automatic concept identification (keyword extraction based on proximity, frequency and other grammatical parameters).
3. Thesaurus learning (the extent to which collections of concepts ‘travel together’ through the text is quantified and clusters formed).
4. Concept location (blocks of text are tagged with the names of concepts which they may contain).
5. Mapping (a visual representation of the semantic network is produced showing how concepts link to each other).
6. Network analysis (this stage is not a part of the Leximancer™ package but was carried out as an additional step to define the structural properties of the semantic networks).

3 Results and Discussion

3.1 Semantic Extraction

A metric for the amount of semantic content able to be extracted from different road scenarios is given by the word count of the verbal transcripts. The total word count across all road types and both road users is 28,169. Under the null hypothesis the total word counts for motorcyclists and car drivers should be 14,084 (i.e. 28,169 / 2). In fact the findings show the total word count for motorcyclists (16,678) to be 18% higher than that for car drivers (11,491). This occurs despite motorcyclists spending on average approximately 3 minutes less time traveling around the course. The largest difference in word count occurs in motorway driving and junctions, with motorcyclists providing 23% and 20.7% more verbal content respectively than car drivers. Controlling for the effect of each road section's mileage to produce a normalised 'words per mile' metric reveals a distinct pattern. Overall, the fastest roads, with speed limits of 70 mph (i.e. motorways), 60 mph (i.e. major and country roads) and 40 mph (i.e. urban roads) produce less than 150 words per mile. Junctions and residential roads (with 30 mph limits) produce in excess of 350. The first point to make is a methodological one. Clearly there is sufficient spare mental capacity, particularly for motorcyclists, for a rich verbal commentary to be provided across all road types. Indeed, the more challenging road types yield more content rather than less, which is what interference due to workload might otherwise suggest. The second point is a theoretical one. It is evident that motorcyclists are able to extract more semantic content from the same situations than car drivers. Furthermore, it would seem that the quantity of semantic content is contingent on the speed and hazard incident rate of different road types. Hazard incident rate is a concept used in police driver training. A hazard is defined by Coyne (2000) as anything potentially dangerous and/or has the potential to cause the driver to change the position and/or speed of their vehicle. Clearly, some road types such as motorways, with restricted access, grade separated junctions, lower speed differentials and gentle alignments have a lower hazard incident rate than a busy urban road, with unrestricted access, at-grade crossings and potentially unfavourable geometry. In other words, 30 mph in an urban setting typically provides many more hazards per mile than 70 mph on a motorway. Differences in word count, therefore, seem to reflect the presence of more stimuli. Whether more stimuli might lead to deeper and/or different reasoning entirely is the topic of the next sections.

3.2 Structure of Semantic Networks

A total of 12 semantic networks are produced from the semantic content captured in the verbal transcripts, six for each of the two road user types (motorcyclist and car driver). These six networks refer in turn to the six road types encountered around the test route (motorway, major, country, urban, residential roads and junctions).

Analysis of these networks now proceeds on the basis of their structure. The structural analysis employs techniques from graph theory to view the semantic networks in terms of nodes (n) and edges (e). These procedures help to reveal important underlying structural properties of the semantic networks which are not readily apparent from visual inspection alone. The metrics used are: density, diameter and centrality.

Density is given by the formula:

$$\text{Network Density} = \frac{2e}{n(n-1)} \tag{1}$$

where e represents the number of edges or links in the semantic network and n is the number of nodes or semantic concepts. The value of network density ranges from 0 (no concepts connected to any other concepts) to 2 (every concept connected to every other concept; Kakimoto et al., 2006). Density is a metric which refers to the semantic network as a whole and is a measure of its overall level of interconnectivity.

Higher levels of interconnectivity suggest a richer set of semantic links and a well integrated set of concepts. A more dense network is also likely to have more well connected concepts and shorter average path lengths. In order to diagnose the latter, a further metric is employed: diameter.

Diameter is given by the formula:

$$\text{Diameter} = \text{maxuyd}(\text{ni, nj}) \tag{2}$$

where d(ni, nj) is "the largest number of [concepts] which must be traversed in order to travel from one [concept] to another when paths which backtrack, detour, or loop are excluded from consideration" (Weisstein, 2008; Harary, 1994). Diameter, like density, is another metric which refers to the network as whole. Generally speaking, the bigger the diameter the more concepts within the semantic network that exist on a particular route through it. Again, generally speaking, a more dense network will have smaller diameter (because the routes across the network are shorter and more direct) while a less dense network will have a larger diameter (as routes across the network have to traverse a number of intervening semantic concepts). This measure is related to the idea of clustering and to individual semantic concepts which are more or less well connected than other concepts. In order to diagnose this facet a further metric is deployed: centrality.

Centrality is given by the formula:

$$\text{Centrality} = \frac{\sum_{i=1;j=1}^{g} \delta ij}{\sum_{j=1}^{g} (\delta ij + \delta ji)} \tag{3}$$

where g is the number of concepts in the semantic network (its size) and δji is the number of edges (e) on the shortest path between concepts i and j (or geodesic distance; Houghton et al., 2006). Centrality gives an indication of the prominence that each concept has within the semantic network. Concepts with high centrality have, on average, a short distance (measured in edges) to other concepts, are likely to be well clustered and to be near the centre of the network. Concepts with low centrality are likely to be on the periphery of the network and to be semantically distant from other concepts.

The mean level of interconnectedness of the semantic networks (as measured using the density metric) is 0.07 for car drivers and 0.08 for motorcyclists. This difference is very small as demonstrated by the almost identical level of density across most road types. However, it can be observed that the semantic networks for motorcyclists becomes more densely interconnected when travelling over residential roads (0.12 compared to 0.08).

The results for diameter show that whilst the overall level of interconnectedness is broadly similar across road user types, as road speeds, and hazard incident rates, increase, the diameter of the semantic networks for motorcyclists decreases. This means that the extent of direct access to semantic concepts increases with hazard incident rate. The reverse trend seems true for car drivers.

Analysis of the metric centrality shows that, overall, as road speeds decrease so too does the mean level of clustering. In other words, as speeds increase specific semantic concepts become much more relevant than others. An exception to this overall pattern is when travelling over country roads, where the average level of clustering increases markedly for car drivers. A less dramatic increase was also observed for both road users in respect to junctions where the mean level of clustering increases once more.

In summary, the overall level of semantic interconnectivity is broadly comparable between motorcyclists and car drivers. The main finding is that while word counts increase with hazard incident rate, the prominence of individual concepts tends to decrease (for both road users). Another key structural difference between motorcyclists and car drivers seems to be in respect to diameter, whereby average path lengths between semantic concepts decrease with hazard incident rate for motorcyclists (suggesting a more integrated mental representation), and increase for car drivers (suggesting a less integrated mental representation).

3.3 Thematic Analysis

In Leximancer™ concept groupings are referred to as 'themes'. These help to raise the level of analysis from the individual items of sometimes rather idiosyncratic keywords to that of broader, highly connected clusters related to how a situation is interpreted. Themes are ascribed a relevance value by Leximancer™. This is derived from the number of times the theme occurs as a proportion of the most frequently occurring concept (Smith, 2003).

There are a total of 64 individual themes extracted from the 12 semantic networks. Not all of these themes score highly in terms of relevance so the data is filtered in order to capture those scoring in excess of 70% within either (or both) the motorcyclist and/or car driver data sets. The filtering process reduces the number of themes from 64 to a high scoring subset of 20. Table 1 presents a summary of the results obtained. Under the null hypothesis it would once more be expected that the matrix of populated cells and the relevance values they contain, would be the same for motorcyclists and car drivers (a difference value of zero). Once more, this is not the case.

Table 1. Summary of results

	Motorway	Major Road	Country Road	Urban Road	Residential Road	Junction
Number of themes increasing in relevance for Motorcyclists	6	4	4	2	1	4
Number of themes increasing in relevance for Car Drivers	5	2	5	4	4	4
Number of themes remaining the same for both road users	9 (45% Overlap)	14 (70% Overlap)	11 (55% Overlap)	14 (70% Overlap)	10 (50% Overlap)	14 (70% Overlap)

Visual inspection of Table 1 reveals differences between road users. Out of the 120 cells contained in the matrix, 49 are not equal to zero. Of those 49, 21 show differences in relevance of 70% or more. Of those 21, 11 have increased relevance for motorcyclists and 10 for car drivers. In summary, then, there is 59.2% thematic overlap between motorcyclists and car drivers but 41.8% of strong thematic difference. This overall difference continues into road types. For motorcyclists, the pattern of results is consistent with the earlier findings on network diameter. Generally speaking, as road speed decreases and the hazard incident rate increases, the number of themes, and their relevance, tends to drop. This finding appears to triangulate with the findings presented above on centrality.

4 Conclusion

This short paper has tried to show how a reliable, automated, semantic network creation process, coupled to concurrent verbal protocol data, is able to provide some interesting insights into how different road users experience the same road situations. From this analysis it is clear that motorcyclists interpret the same road situations differently to car drivers. In many road circumstances this interpretation appears to have important areas of mutual reinforcement, with strong stereotypical responses which favour the anticipation of each other. This is not the case for all road types. Not surprisingly, the two road types of most concern to motorcyclists (junctions and country roads) are interpreted differently and in ways that are more difficult to reconcile for both road users. The exploratory analysis described in this study is compatible with a number of more practical accident analysis and prevention measures, all of which present themselves as candidates for further in depth study.

The use of verbal protocols, task talk-throughs, interviews and focus groups is well established in the human factors literature as a way of defining information and training needs. The present analysis method could help to define such needs in the realm of driving, helping to equip drivers with a form of 'meta-awareness' of their own propensity towards certain cognitive states in certain situations. For example, driver training could provide coaching and tuition on the need to conduct regular rearward checks on country roads and at junctions (as faster vehicles may be approaching from behind). Infrastructural interventions suggested by this work could involve road signs that do not warn of events ahead, but instead warn of potential events behind (e.g. 'faster traffic approaching behind', 'check mirrors now' etc.).

A further practical intervention is the concept of cross-mode training. This already represents best practice in several transport domains. For road transport there is distinction to be made between specific vehicle control skills (e.g. clutch control, hill starts, reversing etc) and mode-independent skills (e.g. road and traffic awareness, giving indications, rights of way etc.). Interventions of this form could take the form of practical training of the latter mode-specific skills using alternative vehicle types, simulations, walk and or talk-throughs. The aim would be to provide practical experience of how different road users interpret the same situation.

The final intervention suggested by the present work relates again to the railway industry and the concept of 'route drivability' (Hamilton, Lowe & Hill, 2007). In essence, this is a form of 'analytical prototyping' in which proposed changes in routes, signaling, signage etc. are tested in terms of driver workload. For road transport, the verbal protocol/semantic network method (in cooperation with other methods) could serve a similar analytical prototyping purpose. The method outputs could be used to ascertain how road situations are interpreted, how physical features could be used to modify that interpretation in favourable ways, and to define cognitively the optimum type and placement of road signs and other infrastructure.

References

1. Anderson, J.: The architecture of cognition. Harvard University Press, Cambridge, MA (1983)
2. Ausubel, D.: The psychology of meaningful verbal learning. Grune & Stratton, New York (1963)
3. Bartlett, F.C.: Remembering: A study in experimental and social psychology. Cambridge University Press, Cambridge (1932)
4. Crandall, B., Klein, G., Hoffman, R.: Working minds: A practitioner's guide to cognitive task analysis. MIT Press, Cambridge, MA (2006)
5. Eysenck, M.W., Keane, M.T.: Cognitive psychology. Laurence Earlbaum, Hove (1990)
6. Gibson, J.J.: The ecological approach to visual perception. Houghton Mifflin, Boston (1979)
7. Hamilton, I.W., Lowe, E., Hill, C.: Early route drivability assessment in support of railway investment. In: Wilson, R., Norris, B., Clarke, T., Mills, A. (eds.) People and rail systems: human factors at the heart of the railway, Ashgate, Aldershot (2007)
8. Harary, F.: Graph Theory. Addison-Wesley, Reading, MA (1994)
9. Houghton, R.J., Baber, C., McMaster, R., Stanton, N.A., Salmon, P., Stewart, R., Walker, G.H.: Command and control in emergency services operations: a social network analysis. Ergonomics 49(12-13), 1204–1225 (2006)
10. Kakimoto, T., Kamei, Y., Ohira, M., Matsumoto, K.: Social network analysis on communications for knowledge collaboration in OSS communities. In: Proc. The 2nd International Workshop on Supporting Knowledge Collaboration in Software Development (KCSD 2006), Tokyo, Japan, pp. 35–41 (2006)
11. Neisser, U.: Cognition and reality: principles and implications of cognitive psychology. Freeman, San Francisco (1976)
12. Norman, D.A.: The design of everyday things. Doubleday, New York (1990)
13. Salmon, P.M., Stanton, N.A., Walker, G.H., Jenkins, D.P.: Distributed situation awareness: advances in theory, measurement and application to teamwork. Ashgate, Aldershot (2009)
14. Smith, A.E.: Automatic extraction of semantic networks from text using leximancer. In: Proceedings of HLT-NAACL, Edmonton (May-June 2003)
15. Stanton, N.A., Salmon, P.M., Walker, G.H., Jenkins, D.P.: Genotype and phenotype schemata and their role in distributed situation awareness in collaborative systems. Theoretical Issues in Ergonomics Science, 10(1), 43–68 (2009)
16. Weisstein, E.W.: Graph Diameter. From MathWorld–A Wolfram Web Resource (September 17, 2008), http://mathworld.wolfram.com/GraphDiameter.html

Part III
Cognition and the Web

Information Searching on the Web: The Cognitive Difficulties Experienced by Older Users in Modifying Unsuccessful Information Searches

Aline Chevalier[1], Aurélie Dommes[2], and Jean-Claude Marquié[1]

[1] Laboratoire CLLE-LTC (UMR 5263, CNRS, Université de Toulouse, EPHE), Maison De la Recherche, 5 Allées Machado, 31058 Toulouse Cedex 9, France
[2] IFSTTAR, French institute of science and technology for transport, development and networks. 25 allée des Marronniers, 78 000 Versailles-Cedex, France
{aline.chevalier,marquie}@univ-tlse2.fr,
aurelie.dommes@ifsttar.fr

Abstract. The present study addressed age-related differences in performances and strategies developed by web users while searching for information. Ten older and 10 younger adults had to search for information with Google and to answer 9 questions varying in complexity: from simple ones (participants needed to use keywords provided in the questions) to impossible ones (no answer existed). The results showed that older participants had lower performances than younger ones; age-related differences were more particularly marked as the question complexity increased. Regression analyses showed that processing speed and cognitive flexibility accounted for a large part of the variance in performances. The younger and older participants also differed in the strategies they developed while searching for information. The older participants tended to focus on the evaluation of the results provided by Google. In contrast, the younger participants tended to plan and regulate their activity, this last strategy provided better performances.

Keywords: Information searching; aging; question complexity; cognitive abilities; strategies.

1 Introduction

The information searching (IS) activity can be considered as problem-solving and decision-making activities whereby the problem-solver's knowledge and other mental representations are manipulated in order to achieve a goal. More precisely, according to Sharit, Hernandez, Czaja and Pirolli [11], the information problem-solving process is divided into the three following sub-processes:

- The representation of the problem to be solved. The problem statement is internalized in order to build up a mental representation of the information elements to be searched.
- The planning. A method for coming up with a solution is elaborated. It often requires dividing the problem into sub-goals.
- The execution. The operations that were elaborated during the planning process are carried out.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 225–232, 2011.

These three processes are iterative as planning may generate further insights into the problem and thus promotes modified problem representations.

IS using a search engine requires more particularly the individuals to generate keywords relevant to their request, to evaluate the relevance of the results provided by the search engine, and then to select one or more web pages to be visited. If the search engine does not provide the expected result, the IS activity becomes more complex: the individuals have to reformulate their first request by adding and/or suppressing keywords. Reformulating unsuccessful requests may be a high-demanding task and involves several cognitive abilities and processes such as processing speed, flexibility and working memory. Processing speed and working memory may allow individuals to remember information previously seen and to compare it to the new elements displayed on the current webpage. Cognitive flexibility may be involved to modify search strategy and requests, and vocabulary abilities to generate new keywords.

Cognitive flexibility — defined as the ability to switch between cognitive strategies in order to adapt to unexpected conditions in the environment [1, 6] — working memory and processing speed are usually reported to decline with aging. In contrast, vocabulary is commonly reported to be higher in older adults than in their younger counterparts.

In a recent study comparing younger and older web users, we showed that cognitive flexibility was strongly involved in the reformulation of unsuccessful requests and that age-related decline in flexibility ability affected searching performances [4]. Based on this previous work, we carried out a new experiment with younger and older web users. The aim was to replicate earlier findings and further identify the cognitive processes involved in IS and that may explained age-related differences in performances and strategies. Based on the models proposed in the literature [11, 14], the IS activity of the younger and older web users has been analyzed according to four activity components:

- Planning: individuals elaborate a plan to reach IS goal. The plan allows them to formalize the first request.
- Evaluating the relevance of the information displayed by the search engine or in the websites by comparison with the information question.
- Regulating: if the first request does not provide relevant result(s), individuals have to modify their strategy to find the answer.
- Monitoring: individuals continuously recall the information problem at hand and examine whether the actions made allow them to achieve the search goal.

2 Method

2.1 Participants

10 younger adults (age range: 21-27 years; M=24.6, SD=1.78) and 10 older adults (age range: 60-68 years; M=62.6, SD=2.12) volunteered to participate to the study. All participants were French native speakers, in good health, and had normal or

corrected-to-normal vision. Education ($M_{younger}$=15.1 years, $SD_{younger}$=1.52; M_{older}=15.2 years, SD_{older}=1.87) and familiarity with Internet were controlled (Internet was used for at least five years for information searching, mailing, and chatting).

2.2 Experimental Procedure

The participants first took a battery of cognitive tests:

- The letter comparison test [10] was used to measure processing speed. The participants had 30 sec. to examine as many pairs of letters (e.g., X O) as possible and decide whether the two letters of the pair were similar or different.
- The Corsi block test (initially developed by [9]) was used to assess working memory [12]. A 4×4-cells grid was shown to participants on a computer screen. Visual patterns were formed by the successive blackening of a variable number of cells at each trial. Immediately after each presentation, the participants had to reproduce the sequence in the correct order by moving the finger on this interface.
- The Trail Making Test reflected cognitive flexibility abilities (part B) [13]. Twenty five circles were distributed over a sheet of paper; the circles included both numbers (1 – 13) and letters (A – L). Participants were asked to draw lines as quickly as possible to connect the circles in an ascending pattern, and alternating between the numbers and letters (i.e., 1-A-2-B-3-C, etc.).
- The French version of the Raven Mill Hill Vocabulary scale was used to assess vocabulary skills [3]. Part 1 asked the participant to define 44 words. The part 2 consisted of a series of 44 words. For each word the participant had to choose a synonym from a list of 6 words.

Then, by using Google, the participants had to answer 9 information questions varying in complexity:

- 3 simple questions: the keywords required to obtain the right answer were included in the statement of the question.
- 3 difficult questions: the relevant keywords needed to obtain the correct answer were not included in the search statement. The participants had to generate the accurate keywords by themselves.
- 3 impossible questions: these questions contained elements that led participants to believe that an answer could be found. Actually no answer existed on the Web.

While performing the IS activity, participants were asked to think aloud [5] in order to identify the strategies they carried out.

2.3 Variables and Data Analyses

The effect of age on cognitive abilities was examined using t tests. The following performances were considered:

(1) Processing speed, was reflected by the number of correct ratings ("Same" or "Different" response) on the letter comparison test.

(2) The working memory span corresponded to the length of the highest correct sequence that the participant was able to reproduce in the Corsi block test.

(3) Cognitive flexibility ability was reflected by the time taken by the participant to complete the Trail Making Test.

(4) A vocabulary score was computed (maximum score = 88 points) from responses of each participant to the Mill Hill Test.

Three dependent variables were selected with regard to the information search task:

(1) Search times (in sec.). For each experimental search question, the search time was calculated from the moment the participants started the search with Google (i.e. after keywords were entered in the text box) until they highlighted the answer by clicking the mouse, or until they declared abandoning the attempt to find the answer. The keywords typing time was separately computed and removed from the search time. Therefore, this measure corresponded to the total time taken by the participant to answer a search question, including all formulated requests and excluding typing speed of all keywords entered in the text box.

(2) Number of correct answers. For simple and difficult search questions, a correct answer was scored 1. An incorrect answer or abandoning the attempt to find the answer was scored 0. For impossible questions, both abandoning and the participants' indication that no answer did exist were scored 1.

(3) Number of reformulations of the first request. For each search question, we computed the number of request reformulations that the participant formulated and entered in the search engine text box after a first unsuccessful request.

These three dependent measures from the IS task were input into analyses of variance (ANOVAs) with age (young, old) as a between-group factor, and search question complexity (simple, difficult, impossible questions) as an intra-group factor. Qualitative analyzes were also conducted on the basis of the participants' recorded verbal protocols, especially the planning, evaluating, regulating and monitoring activities while older and younger participants searched for information on Internet.

3 Results

3.1 Cognitive Abilities

- Processing speed

Young adults (M=30; SD=3.13) showed higher processing speed scores than the older ones (M=21.5; SD=3,41) (t(18)=5.812, p<.0001).

- Working memory

Young adults (M=6.7, SD=1.25) showed greater working memory spans than the older ones (M=5.7, SD=0.82) (t(18)=2.111, p<.05).

- Cognitive flexibility

The time required to solve the part B of the Trail Making Test was affected by age (t(18)= 2.474, $p<.05$), with young adults (M=65.32; SD=25.33) taking less time than the older ones (M=102.2; SD=39.75).

- Vocabulary ability (Mill Hill)

Older participants (M= 72.9; SD=5.53) outperformed their younger counterparts (M=67.2; SD=7.07) (t=-2.009, p=.05) at the vocabulary test.

3.2 Information Searching Performances (see Table 1)

- Number of correct answers

ANOVAs revealed a significant effect of age (F(1,18)=12.521, $p<.005$, η_p^2=.41) and search complexity (F(2,36)=51.095, $p<.0001$, η_p^2=.74) on the number of correct answers. The older adults provided less correct answers than the younger ones. Simple questions generated more correct answers than the difficult (F(1,18)=110.769, $p<.0001$) and impossible ones (F(1,18)=17.894, $p<.001$). Difficult questions also generated less correct answers than the impossible ones (F(1,18=29.16, $p<.0001$).

The Age × Complexity interaction was also significant (F(2,18)=3.722, $p<.05$, η_p^2=.17). More precisely, planned comparisons showed that the young participants found more correct answers than the older for the difficult (F(1,18)=5.651, $p<.05$) and impossible questions (F(1,18)=8.228, $p<.05$)

- Search times (in sec.)

Age showed no significant effect on search times (F(1,18)=0.058, $p>.1$). In contrast, we observed a significant effect of search complexity (F(2,36)= 47.3782, $p<.00001$, η_p^2 =0.73). More precisely, planned comparisons showed that the time to answer simple questions was smaller than the time needed to answer impossible questions (F(1,18)=72.06, $p<.00001$) and difficult ones (F(1,18)=68.94, $p<.00001$). The impossible questions were solved quicker than the difficult ones (F(1,18)=8.41, $p<.01$).

The Age x Complexity interaction was not significant (F(2,36)=1.3505, $p>.1$).

- Number of reformulations of the first request (and the next ones)

The young participants reformulated more often their unsuccessful requests than the older ones (F(1,18=12.019, $p<.01$, η_p^2=.40). The search complexity also had a significant effect (F(2,36)=22.383, $p<.00001$, η_p^2 =0.55). The planned comparisons showed that simple questions generated less reformulations than the impossible (F(1,18)=43.30, $p<.00001$) and difficult ones (F(1,18)=31.56, $p<.0001$). No significant difference was observed between the difficult and impossible questions (F(1,18)=0.55, $p>.1$).

The Age × Complexity interaction was significant (F(2,36)=3.819, $p<.05$, η_p^2=.18). More precisely, the young participants reformulated more often unsuccessful requests than the older participants only when dealing with difficult questions (F(1,18)=5.895, $p<.05$) and impossible questions (F(1,18)=10.069, $p<.001$).

Table 1. Mean (SD) values for the number of correct answers (3-point scale), search time (in sec.) and the number of reformulations of unsuccessful request according to age group and complexity of the search

Dependant variables	Age groups	Simple questions		Difficult questions		Impossible questions		*Total*	
		M	SD	M	SD	M	SD	*M*	*SD*
Correct answers	Young	2.9	0.32	1.4	0.7	2.7	0.48	*2.15*	*0.41*
	Older	3	0	0.5	0.97	1.9	0.74	*2*	*0.49*
	Total	*2.95*	*0.22*	*0.95*	*0.94*	*2.3*	*0.73*	*1.95*	*0.48*
Search time	Young	48.13	22.44	361.33	207.58	310.27	162.67	*239.91*	*203.01*
	Older	52.07	50.92	433.07	151.84	262.63	45.49	*249.26*	*183.64*
	Total	*50.1*	*38.35*	*397.2*	*180.79*	*286.45*	*118.79*	*244.58*	*84.14*
Reformulations	Young	0.27	0.26	4.77	3.12	4.33	2.35	*3.12*	*1.48*
	Older	0.1	0.22	2.07	1.62	1.67	1.25	*1.28*	*0.79*
	Total	*0.18*	*0.25*	*3.42*	*2.79*	*3*	*2.28*	*2.2*	*1.5*

3.3 The Role of Age and Cognitive Abilities in Reformulating Unsuccessful Requests

A multiple regression analysis was performed to investigate the respective roles of age and cognitive abilities in the number of request reformulations made to achieve the goal of answering the 9 search questions. Age, processing speed, working memory span, cognitive flexibility and vocabulary skill were used as predictors.

The results showed that processing speed (β=1.06; p<.01) and cognitive flexibility (β=0.66; p<.05) both emerged as significant predictors accounting for 71% of the variance in reformulations.

Working memory and vocabulary scores were not significant and did not account for a significant amount in the variance explained by the model. Age was not a significant predictor of the variance in reformulations.

3.4 Distribution of the IS Activity Components

Qualitative analyses of verbalizations showed differences in the distribution of planning, evaluating, regulating and monitoring activities as a function of age.

The young participants tended to regulate and plan more often their IS activities than evaluate the search results. By contrast, the older participants evaluated the search results more often than they regulated and planned their activity.

Monitoring strategies were very little reported by participants.

4 Conclusion

The findings of the present study showed poorer information searching performances in older participants. Although both the young and older participants took about the same time to find the correct answers, the young participants performed better than the older ones.

Age-related differences were higher as task complexity increased. Indeed, older participants experienced specific difficulties in getting out of impasses and reformulating unsuccessful requests. As observed earlier [8], considerable age differences in strategy selection were found in the current study, with older adults tending to rely on simpler strategies minimizing cognitive effort.

However, aging is not the unique factor that affects information searching performances. As underlined by [7], great variability has been observed among older people (for a review see [15]). Some authors reported that age effects are strongly mediated by other factors such as cognitive abilities [2, 11]. In the current study, processing speed and cognitive flexibility were found to explain 71% of the variance in the number of reformulation of unsuccessful requests. Age was not a significant predictor once the user's processing speed and cognitive flexibility abilities were taken into account.

Our findings are in line with those obtained in a previous study showing a strong influence of cognitive flexibility on information searching performances, especially as task complexity increased and required a change of strategies [4]. Therefore, older adults may experience difficulties in getting out of impasses and reformulating their requests because of the decline of their cognitive flexibility abilities and the slowing of their processing speed. Their higher vocabulary skills did not seem to counteract these cognitive declines.

The qualitative analyses of verbalizations, recorded while the participant was performing the information searching task, pointed out age-related differences for difficult and impossible questions in some task components such as planning, evaluating and regulating. Young participants were shown to report more planning and regulating activities than evaluation of the pertinence of the results provided by the search engine. In contrast, older participants tended to evaluate more often the relevance of the results provided by Google than planning and regulating their searching activities. Older participants appeared to experience difficulties in modifying unsuccessful strategies compared to their younger counterparts. This may reflect less confidence in their searching strategies.

These last findings suggest that the difficulties experienced by older adults in getting out of impasses are related to difficulties in planning their searching activity as well as to difficulties in modifying their strategies for better ones. Even though they tried to compensate for these difficulties by a high number of evaluations, older adults may have experienced more difficulties in producing efficient strategies.

To conclude, the present study underpins two relevant points:

1) The relevance of training methods well suited to the specific needs of older users and that may help them to develop a more efficient searching activity.
2) The importance of designing search engine tools that take into account the difficulties of older users as well as the age-related declines in cognitive abilities.

References

1. Chevalier, A., Chevalier, N.: Influence of proficiency level and constraints on viewpoint switching: A study in web design. Applied Cognitive Psychology 23, 126–137 (2009)
2. Czaja, S.J., Sharit, J., Ownby, R., Roth, D., Nair, S.: Examining age differences in performance in a complex information search and retrieval task. Psychology and Aging 16, 564–579 (2001)
3. Deltour, J.J.: Échelle de vocabulaire Mill Hill de J.-C. Raven, French Version. Éditions et Applications Psychologiques, Paris (1998)
4. Dommes, A., Chevalier, A., Lia, S.: The Role of Cognitive Flexibility and Vocabulary Abilities of Younger and Older Users in Searching for Information on the Web. Applied Cognitive Psychology (in press)
5. Ericsson, K.A., Simon, H.A.: Protocol analysis: Verbal reports as data (Revised edition). MIT Press, Cambridge (1993)
6. Eslinger, P.J., Grattan, L.M.: Frontal lobe and frontal-striatal substrates for different forms of human cognitive flexibility. Neuropsychologia 31, 17–28 (1993)
7. Lindberg, T., Näsänen, R., Müller, K.: How age affects the speed of perception of computer icons. Displays 27, 170–177 (2006)
8. Mata, R., Nunes, L.: When less is enough: Cognitive aging, information search, and decision quality in consumer choice. Psychology and Aging 25, 289–298 (2010)
9. Milner, B.: Interhemispheric differences in the localization of psychological processes in man. British Medical Bulletin 27, 272–277 (1971)
10. Salthouse, T.A.: Working memory as a processing resource in cognitive aging. Developmental Review 10, 101–124 (1990)
11. Sharit, J., Hernandez, M.A., Czaja, S.J., Pirolli, P.L.: Investigating the roles of knowledge and cognitive abilities in older adult information seeking on the Web. ACM Transactions on Computer-Human Interaction 15, article 3 (2008)
12. Smyth, M.M., Scholey, K.A.: Determining spatial span: The role of movement time and articulation rate. The Quarterly Journal of Experimental Psychology 45, 479–501 (1992)
13. Spreen, O., Strauss, E.: A compendium of neuropsychological tests: Administration, norms, and commentary. Oxford University Press, New York (1991)
14. Tricot, A., Rouet, J.-F.: Activités de navigation dans les systèmes d'information. In: Hoc, J.-M., Darses, F. (eds.) Psychologie ergonomique: tendances actuelles, pp. 71–95. PUF, Paris (2004)
15. Wagner, N., Hassanein, K., Head, M.: Computer Use by Older Adults: A Multidisciplinary Review. Computers in Human Behavior 26, 870–882 (2010)

Template for Website Browsing

Fong-Ling Fu[1] and Chiu Hung Su[2]

[1] Department of Management Information Systems National Chengchi University, No. 64, Sec. 2., ZhiNan Rd., Wenshan District, Taipei City 11605, Taiwan (R.O.C.)
flfu@nccu.edu.tw
[2] Department of Management Information Systems Hwa Hsia Institute of Technology
ritasu@cc.hwh.edu.tw

Abstract. Websites on e-commerce often display large amounts of multi-media and information, creating problems for viewers when locating specific information. This research uses the concepts of template and selective attention to understand the cognitive simplification in finding information and browsing websites. Utilizing content analysis with 240 university students as subjects, we conducted an experiment on information retention with browsing a shopping website. Although the amount of information displayed by the website was staggering, the result of the experiment showed that participants applied a template built up through past experiences of what's important and where things belong. This internal map containing three mechanisms: segmentation, grouping and attention, is then used to create an efficient task strategy, to segment the page, and to categorize the information. This research tried to understand the attributes of template for users who are browsing Websites. The "findability" of online information would be improved if the arrangement of information of a web site were the same as what viewers expected.

Keywords: Information search ability, selective attention, interface design of websites, template matching.

1 Introduction

How to improve searching ability is a hot issue in website design now, because the continuous trend in websites to display ever increasing amounts of textual information and multimedia content is resulting in information overload that is not only visually confusing, but complicates information searching [9, 8]. The reason why information is considered complicated is because sometimes there is too much of it to be processed by the brain. When the information is first read in by the human eye, hundreds of thousands of optical neurons are in place to register vast amounts of data. In fact, as understood by the mental perceptual model, the impressive processing capability of the optical nerves is able to handle virtually any volume of information elements [17].

The brain however, has only seven chunks in its short term memory, and this limited capacity is unable to fully process everything relayed from the eyes into meaningful and useful information [12]. To cope with this mismatch in processing ability and input volume, the mind uses a filter mechanism to sort the input, by only

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 233–242, 2011.

selectively processing information understood to be interesting or important, and ignoring information irrelevant to the user's task at hand; this is commonly referred to as "selective attention". Therefore when a viewer looks at complex information, his or her eyes will register all the objects, but his or her mind will not "see" and use all the details; the ignored messages will not be mentally processed any further [12, 1].

For more understanding the human being's mechanism of information filter, we hypothesize that when viewers see a webpage, they use what cognitive theory calls "template matching", to apply a template of what's important and where things belong, built out of past experiences. This internal map is then used to create an efficient task strategy, to segment the page, and to categorize the information [12, 13]. This experiment sets out to find what the template probably is, with the hopes that the findings may lead to identifying what types of attribute variety is needed for clear distinction of information and which ways of grouping information are the most suitable to particular kinds of tasks. Ultimately we hope to improve the way websites are designed and making them more accessible and useable for viewers.

2 Template-Matching: The Mechanism to Reduce Complexity

Template-matching Technique was proposed by Ben and Funder [2] which provided a language of description for both persons and situations. Applied to prediction of human being's behavior, the technique consists of two basic steps. First, each conceptually distinct behavior in the laboratory setting under investigation is characterized by a *template,* a personality description *of* the hypothetical idea person who is most likely to display the design behavior in that situation. These templates serve to characterize the potential behavior of subjects inside the laboratory. Second, personality descriptions of subjects are obtained from close acquaintances; these descriptions characterize the subjects' behaviors outside the laboratory. A particular individual behavior was predicted through the match between these two sets of data. It has been used in explaining peoples behaviors in social psychology [3]. Lots of applications of template-matching technique in image recognition or pattern match which involved template matching processes, assumed that various internal representations (templates) of objects were stored in memory, and new stimuli were processed by comparing them with the templates until a match was found. As a webpage can be considered as a complex image containing lots of information [5], the authors thus try to use template matching technique in predict (explain) behavior of web page browsing.

The way our brain runs the selective attention process is it uses the attributes of what we see, to distinguish the interesting from the unimportant. So the important question is, out of all the attributes to choose from, how does the mind decide which to use? The template states that people just grab a similar template to use because they impossible have enough experiences with all objects and the mental processing time for matching is very fast. Therefore, instead of information contents, information structures are usually used in looking for the pattern of template-matching [14]. Two preprocesses occur before the matching, one is a local operation to eliminate unnecessary noise, and the other is a normalizing process to translate the pictures which over large, over small, lean, or wrapping to be standard or normalized stimulus [10].

3 The Experiment

3.1 Procedures

The experiment was a laboratory experiment using two famous commercial websites in Taiwan: Yahoo Shopping center (http://buy.yahoo.com.tw) and PChome Online (http://shopping.pchome.com.tw). The subject was given one of the websites and assigned a task to identify the most optimal Panasonic digital camera and to find important information about it.

The procedures of the experiment are: first, before browsing the target Website, participants were asked to draw what their ideal start page of an e-commerce Website meant to handle the task. Then, browsing time was limited to only ten seconds per webpage, before users were required to make a mouse click to navigate to the next page, in order to force users to filter out irrelative information and get to key information quickly. The screen was then turned off and subjects would draw a mental image of the webpage just viewed before continuing on. This allowed us to figure out subjects' mental templates and their searching and browsing strategy for each web site. All browsing steps were repeated until the final target was reached.

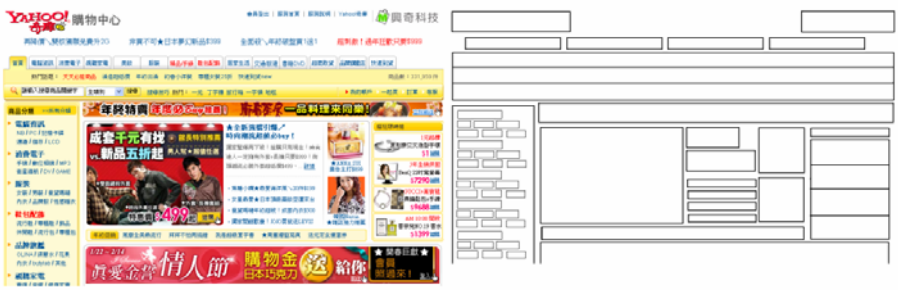


Fig. 1. (a) Yahoo Shopping Center (b) Front page structure model

How can we discover the template used by a viewer during information seeking? Previous studies point to the fact that a viewer can use the structure of information presented to find a pattern [14], and that viewers' cognitive structure for a web site during information seeking thus can be analyzed to find their templates. The method used in the study for drawing structure of a web page was proposed by Ngo et al. [11], to analyze graphical patterns, called structure models. The actual content of the graphics is replaced by squares. For example, Figure 1(a) is a screenshot of the starting page of Yahoo Shopping Center. Removing the actual content turns the page into a structure model, which is illustrated in Figure 1(b).

The actual website is more complex than Figure 1(b), but the structure model is still able to fully capture important details regarding content grouping, location and alignment. When a subject draws out the web pages by memory, they may draw out something far less detailed, as in Figure 2(a), which consists of a frame containing general groupings and rough comments noting why certain chunks were distinguished from others. The subjects' hand-drawings (Figure 2(a)) were then transformed into digital webpage modeling structures, such as the one shown in Figure 2(b). The number of chunks and the information inside the chunks were organized, sorted, and coded through content analyses for further investigation into the subjects' choice of template attributes used and remembered.

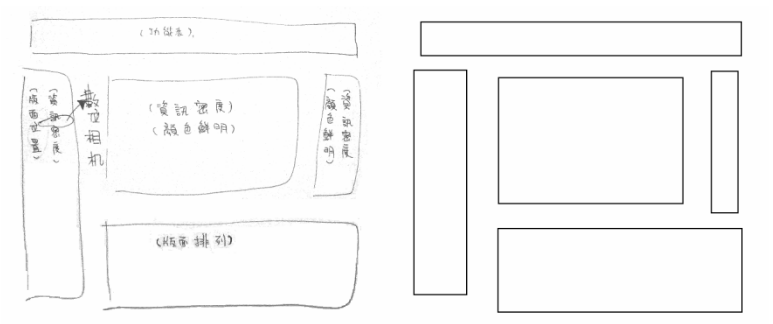


Fig. 2. (a) Subject drawing of webpage (b) Modeling structure of drawing

3.2 Subjects

The subjects were 240 undergraduate volunteers. They all had enough web searching and browsing experience to be adept at using the web. 75% had spent more than 25 hours per week on the Internet, but half of them lacked any online shopping experience. The subjects generally were not familiar with digital cameras. We assume experience with similar digital products would have been adequate for establishing a suitable mental template for information searching even if they were short of actual product knowledge.

To account for differences in subjects' attention capacities and cognition styles, the study utilized an "Embedded Figure Test" to check whether the subjects' cognitive styles were field dependent or field independent [19]. Field independent learners are better at construction, organization, and analysis of information in a multi-media open learning environment than learners who are field dependent [15]. 80% subjects were field independent and 20% were field dependent, but there were no significant differences in results between the two groups in our final analyses.

3.3 Content Analysis

The participants' structure models drawn from short term memory were content analyzed. Content analysis is a method that codes and classifies the qualitative data through frequency distributions. Categories were formed through a systematical, quantitative and objective process. The first step of content analysis was to define the recording unit. Recording unit is the minimal and basic calculation unit used to summarize distribution statistics [18]. Seven recording units were defined in the study to help analyze the amount, varieties, and relations of chunks. These recording units were identified from common themes in participants' notes [6]. The second step of content analyses was to define the categories in each recording unit. Content could only be classified effectively when the categories were defined specifically. Effective categories should be exclusive, exhaustive, and reliable [4, 6].

The subjects have viewed three to four Webpages before ultimately identifying an optimal digital camera. To reduce the amount of content that needs to be analyzed in the process, the study only analyzed the first and last Webpages drawn by the participants. This study argues that the first and last pages were the most critical in understanding an information seeker's mental template. The retained image of the first page hints at a viewer's strategy for information seeking; the chunks remembered illustrate how the individual narrows down available information. The retention picture of the last page visualizes how a viewer's perceives the optimal product.

Our analysis, based on Kalbach's research [7], identified a total of 32 categories that could be utilized to define mental templates. All seven basic units and their respective categories are listed as follows:

1. Use of Hierarchical Searching: whether information searching used the index or relied on browsing the entire page.
2. Number of Chunks Remembered: Further classified into two categories: number of target and non-target (advertisement) chunks retained in short term memory.
3. Number of information Noticed: either 3, 4-7, or more than 7.
4. Reasons for Noticing Target Chunk : included 10 categories: 1) location in familiar position and consistent with experience, 2) key words provided relating to task, 3) significant contrast in colors, 4) distinguishable lines (frame) around the chunk, 5) significant visible alignment among chunks, 6) distinguishable space among chunks, 7) size variety of chunks, 8) located in the noticeable area of screen, 9) font contrast between sections, and 10) distinguishable graph icons. One retention chunk might include multiple above factors.
5. Reasons for Noticing Detail: classified according to the ten categories above.
6. Attention Paid to Target Chunk: included two categories, either Very Concentrated on Target or Viewed Other Information Also.
7. Reason for Action (click): included three categories: attention caught either by graph, by index, or by both.

Because the coded categories were based on the coders' subjective interpretations, reliability became an issue. Three individuals coded the templates independently. Inter-coder reliability was calculated through Hoslti's [6] inter-judge agreement method. Inter-judge agreement scores for the seven basic units of the first Webpage were 0.99, 0.96, 0.95, 0.97, 0.98 and 0.96 respectively, and 0.97, 0.96, 0.95, 0.94, 0.96, and0.93 respectively for last Webpage. The higher liability scores indicate that the assessment process is highly reliable.

Table 1. Comparisons of structure models before and after browsing

Recording Unit	Category	before browsing		After browsing		$\chi 2$	P
		No	%	No	%		
Segment	two areas [1]	102	64	101	63	2.44	0.660
	index only	18	11	18	11		
	two areas[2]	16	10	10	6		
	N/A	20	13	23	14		
	other	4	3	7	4		
Grouping factor	frame	152	95	160	100	9.92	0.042
	font variety	81	51	71	44		
	presence of picture	71	44	99	62		
	contrast in color	38	24	65	41		

Notes: [1] indicated as two areas of index and advertisement.
[2] indicated as two areas of advertisement and other s.

4 Results

No significant difference on p values at Chi-square analyses between the structure models of starting page before and after browsing the Webpage showed that participants had a preconceived expectation of how the start page of a Website is supposed to be structured according the task at hand (Table 1). Based on the task of finding a good digital camera, most participants (64% vs. 63%) noted two segmentations of a web page: one was index area and the other was advertisement area. And the methods which participants grouped chunks included distinguished frame, font variety, presence of picture, and contrast in color.

Table 2 shows the comparisons of frequency with which participants segmented a web page: Most participants segmented a web page as two areas of index and advertisement. But in the initial stage, five percents participants only noted one area of index and in the final page eight percents participants only noted one area of advertisement.

Table 2. Segmentation by location

Category	Starting page		Final page		$\chi 2$	p
	No	%	No	%		
index area only	12	5	0	0		
two areas	226	94	218	91	30.1	0.000**
Adver. Area only	0	0	18	8		

Location seemed to be the most important reason they segment a web page. The segmentation mechanism separated the location of target information in both the index and advertisement area. Participants took almost 3 to 4 “clicks” from the starting page before they found an acceptable product. All participants focused on the index area of the starting page, but their focus changed to the product advertisement area in the final page. The participants’ structure models indicates that their searching strategy involved using the Webpage index to reach the target gradually, first by clicking “digital camera” or “3C” section in the start page and then "Panasonic” on the second page. On the final page, participants ignored the entire index area, and instead paid attention to the product advertisement area.

Table 3. Grouping factors

Category	Starting page		Final page		χ2	p
	No	%	No	%		
distinguished frame	232	97	220	92	16.61	0.02
presence of picture	150	63	175	73		
font variety	114	48	96	40		
containing keywords	107	45	88	37		
contrast in color	89	37	47	20		
size variety	19	8	17	7		

The grouping mechanism breaks down a Webpage into basic units. Participants processed chunks of Web page information according to grouping mechanisms rather than reading everything line by line. Within the area viewed, users would try to distinguish chunks that correspond to their current purpose. As Table 3, the important factors for distinguishing chunks on a web page included the presence of a distinguishable frame and/or margin (97% on structure models of start page and 92% of final page), presence of pictures, (63% of start page and 73% of final page), variety in font (48% on start page and 40% of final page), containing keywords (45% on start page and 37% of final page),color contrast of the chunk from its neighbors (37% of start page and 20% of final page), and variety in chunk size(8% of start page and 7% of final page).

The attraction mechanism that draws viewer’s attention through key words or pictures related to the current task (Table 4, 87% of start page and 98% of final page). Viewers searched for icons with keywords such as “3C” or “digital camera” on the starting page. On the final page containing many different Panasonic camera choices, subjects were attracted to the keyword or image of the product, with particular attention paid to the product’s appearance, function, and price. In both the starting page and the target page, the viewers largely ignored information irrelevant to their current purpose.

Table 4. Attraction with click

Category	Starting page		Final page		$\chi2$	p
	No	%	No	%		
containing keywords	209	87	234	98	79.5	0.000**
contrast in color	2	1	3	1		
distinguished frame	20	8	1	0		
font variety	0	0	1	0		
In first screen	0	0	3	1		
presence of picture	0	0	5	2		

Note: ** indicated as significant at .001.

5 Conclusions

Webpages contain massive, complex information. Understanding how users find their target product can help practitioners create more effective Webpage designs. This study utilized the lens of template and selective attention to assess how viewers would behave when presented with the complex information on a Webpage.

This study conducted a content analysis to verify whether users utilized their cognitive templates formed from past experience to filter information irrelevant to the given task. The result indicated a pattern in their browsing behavior. Through the content analyses of the structure models of information retention, it was then found that three attributes are critical to the viewer template model for information searching behavior: (1) the segmentation mechanism separating the location of target information in both the index and content area. (2) The grouping mechanism breaks down a Webpage into basic units. Participants processed chunks of Web page information according to grouping mechanisms rather than reading everything line by line. (3) The attraction mechanism that draws viewer's attention through key words or pictures related to the current task.

In terms of practical implications, we suggest that Website designers match the layout with users' possible cognitive templates. User templates determine the customers' searching strategies, their preferred location for valuable information, and distinguishable factors of target chunks and icons. We found that location plays a critical role in finding the index in the starting page. Meaningful words and phrases become the most critical factor when users search for target chunks and icons. Advertising chunks in the starting page are ignored by viewers whose searching strategy was to reduce the scope of potentially relevant information. However, once viewers find the target Webpage, they become more willing to browse through the adverting chunks. On the target page, pictures or texts with large fonts serve as the key to attract viewer attention.

This study provided an example of how template can be useful in Webpage design when searching for a specific product. Future research should continue to utilize template in the other kinds of searching tasks. For example, how do people without a clear searching target the browse the Web? How do they look for extended information, or search for "all" types of available product?

Acknowledgements. The research is supported by National Science Foundation of Taiwan (NSC 98-2410-H-004-009).

References

1. Abernethy, B.: Visual Search in Spot and Ergonomics: Its relationship to Selective Attention and Performance Expertise. Human Performance 1(4), 205–235 (1988)
2. Bem, J.B., Funder, D.C.: Predicting More of the People More of the Time: Assessing the Personality of Situations. Psychological Review 85(6), 485–501 (1978)
3. Bem, J.B., Lord, C.G.: Template Matching: A Proposal for Probing the Ecological Validity of Experimental Setting in Social Psychology. Journal of Personality and Social Psychology 37(6), 833–846 (1979)
4. Budd, R.W., Thorp, R.K.: Donohew, L. Content Analysis of Communication. The Macmillan Co., New York (1967)
5. Fu, F.-L., Chiu, S.-Y., Su, C.H.: Measuring the Screen Complexity of Web Pages. In: Smith, M.J., Salvendy, G. (eds.) HCII 2007. LNCS, vol. 4558, pp. 720–729. Springer, Heidelberg (2007)
6. Holsti, O.R.: Content analysis for the social sciences and humanities. Addison- Wesley, Mass (1969)
7. Kalbach, J.: I'm Feeling Lucky: The Role of Emotion in Seeking Information on the Web. Journal of the American Society for Information Science and Technology 67(6), 813–818 (2006)
8. Kuhlthau, C.C.: The Role of the Experience in the Information Search Process of an early Career Information Worker: Perceptions of Uncertainty, Complexity, Construction, and Sources. Journal of the American Society for Information Science and Technology 50(5), 399–412 (1999)
9. Morville, P.: The Age of Findability, Boxes and Arrows (April 29, 2002),http://www.boxesandarows.com//archives/the_age_of_findability.php (Retrieved May 27, 2008)
10. Neisser, U.: Cognitive Psychology. Appleton-Century Crofts, New York (1967)
11. Ngo, D.C.L., Teo, L.S., Byrne, J.G.: Modeling Interface Aesthetics. Information Science 152, 25–46 (2003)
12. Norman, D.A., Bobrow, D.G.: On Data-limited and Resource-limited processing. Cognitive Psychology 7, 44–64 (1975)
13. Pi, Y., Shu, H., Liang, T.: The Frame of Cognitive Pattern Recognition. In: Proceedings of the 26th Chinese Control Conference, China (2007)
14. Reed, C.L., Stone, V.E., Grubb, J.D., McGoldrick, J.E.: Turning Configural Processing Upside Down: Part and Whole Body Postures. Journal of Experimental Psychology: Human Perception and Performance 32(1), 73–87 (2006)

15. Stanton, N.A., Baber, C.: The myth of navigation hypertext: How a bandwagon has lost its course. Journal of Educational Multimedia and Hypermedia 3(3/4), 235–249 (1994)
16. Shugen, W.: Framework of Pattern Recognition Model Based on Cognitive Psychology. Geo-spatial Information Science 5(2), 74–78 (2002)
17. Xing, J.: Measures of Information Complexity and the Implications for Automation Design. National Technical Information Service, Springfield, Virginia (2004)
18. Wimmer, R.D., Dominick, J.R.: Mass Media Research: An Introduction. Wadsworth, California (1994)
19. Witkin, H., Oltman, P.K., Raskin, E., Kaerp, S.A.: A manual for the embedded finures tests. Consulting psychologist Press Inc., California (1971)

Mental Models: Have Users' Mental Models of Web Search Engines Improved in the Last Ten Years?

Sifiso Mlilo and Andrew Thatcher

University of the Witwatersrand Psychology, Wits, 2050, South Africa
sifiso.mlilo@gmail.com, Andrew.Thatcher@wits.ac.za

Abstract. This study investigated the accuracy and completeness of mental models users have of Web search engines in the context of a comparison of matched data obtained from samples from 2000 and 2010. The performance measures time, steps and accuracy were assessed along with 17 salient features of Web search engines identified in the study conducted in 2000. The results indicated that the 2010 sample had improved significantly across all performance measures. The two samples did, however, identify an equal number of salient features (N=7). It was clear from the detailed analyses of the salient features though, that that the accuracy and completeness of users' mental models of search engines had demonstrably improved from 2000. So, while users' mental models of Web search engines still remain largely inaccurate and incomplete, their alignment with designer's conceptualisations has improved.

Keywords: Mental models; search engine; time; steps; accuracy; salient features.

1 Introduction

Web search engines (WSEs) are now recognised as the dominant information seeking tool utilised by people using the Web [1][2]. A number of studies interested in the information retrieving methods employed by users have long recognised this fact [2]. As such, there has been a shift from the focus on users' mental models of the Web in general [3] towards an active interest in the kinds of mental models users have of online library databases [2][4] and WSEs more specifically [1][5][6][7][8][9]. A number of studies that have looked at mental models of WSEs have either done so in a rudimentary fashion [5][7][10] or the mental models were inferred indirectly from log-file data on user queries submitted to WSEs [1][11][12][13]. The log-file analysis (e.g. [13]) found that users' query formulations (and re-formulations) indicated a poor understanding of the manner in which WSEs work. Similar results have been found by researchers like Muramatsu and Pratt [7] with query-based analyses in user studies. Muramatsu and Pratt [7] concluded that users' mental models were "naïve and erroneous" (p. 223) and suggested that WSE interface designers make the query transformations more transparent for their users. Efthimiadis and Hendry [11] found a wide variation in users' mental models of WSEs but in general found them to be

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 243–253, 2011.

simple with few users understanding complex WSE concepts such as link analysis and query parsing. Crudge and Johnson [5] provided a fairly detailed summary of 10 users' mental models of WSEs using a repertory grid and a laddering technique. However, they did not provide any value judgement as to the completeness or accuracy of these mental models or their ability to aid the search process. Zhang [9] found that undergraduates' mental models of WSEs also varied in complexity and accuracy with most users showing fairly naïve mental models that were frequently incorrect. Thatcher and Greyling [8], examining a sample of 80 users, concluded that users' mental models of WSEs were incomplete and inaccurate. Users in their sample frequently misunderstood how search terms worked, used search terms inappropriately (e.g. used Boolean logic operators incorrectly or searched using whole phrases), misunderstood relevance feedback, and rarely used different WSEs to verify information. The general sense from these studies is that users' interactions with WSEs indicate a poor understanding of how they work and hence poor mental models.

When an individual interacts with a particular system they form ideas about how that system works. The more they learn about the system the more likely it is that they will figure out how that system works and respond to certain commands, actions or behaviours. In figuring out how the system works individuals are able to anticipate system responses prior to even engaging the system or giving it the relevant commands. The more a user is exposed to a particular system the better the opportunity for mental models to become more accurate and complete, although a lot depends on the type of exposure and the transparency of the system interface. Experience or knowledge of systems in existence prior to direct exposure also helps users in understanding the new systems they are faced with. All other studies of mental models of WSEs have used a cross-sectional design. This study sought to explore whether mental models of WSE systems have, over time, become more aligned with the designers' conceptualisations of how they work. Applying the same methodology used by Thatcher and Greyling [8], with data collected in 2000, this study aimed to determine how time (and by inference, greater experience) has influenced WSE mental model formation over the last decade. Much has changed in the domain of Web searching since data were collected in 2000. The Web now has far more webpages and websites; this means that webspace is more complex but it also means there are more points where information can be found. Connectivity, for most parts of the World, is considerably faster, making it quicker for users to find information without getting frustrated by slow download speeds. Importantly for mental model formation, users are more likely to have been exposed to WSEs for longer periods of time. Finally, the WSE landscape has changed considerably where no single WSE dominated the Web search environment in 2000. Since the exposure people have to the Web is likely to have increased quite significantly since the Thatcher and Greyling [8] investigation, a meaningful platform was set to assess whether the changes since 2000 have seen a closer alignment of users' mental models with designers' conceptual models of WSEs.

2 Method

2.1 Sample

For the 2010 sample a total of 80 students, scholars, professionals and semi-professionals, were selectively targeted for voluntarily participation in the study. This purposive sampling technique was used to attempt to match the 2010 sample as closely as possible to the 80 participants in the 2000 sample. Respondents were matched as closely as possible on age, language, gender, and occupation (e.g. student, scholar and type of occupation). In the 2000 sample there were 50 male participants (48 male respondents in 2010), a mean age of 23.28 years (25.78 years in 2010), 51 respondents who spoke English (50 in 2010), 25 who spoke an African language from South Africa (24 in 2010), and 2 who spoke a language from elsewhere in Africa (5 in 2010). Since the sample in 2000 was anonymous it was not possible to trace the exact same participants which would have been ideal. The 2010 sample must therefore be considered as a contrast group.

2.2 Procedure

Upon their arrival at the location where the data was being collected and following their completion of the biographical and composite WWW experience questionnaire, participants were asked to complete a directed search task while having their search actions recorded by HyperCam3 onscreen capture software. Because there was a possibility that some participants (expert users or subject experts) would not use a WSE for the directed search task, a secondary task (i.e. a general purpose browsing task) requiring them to use a WSE was also put in place. Only two participants each in 2000 and 2010 did not use a WSE for the directed search task and therefore also completed the secondary task. In these two cases the directed search task was used for the performance measure and the general purpose task was used to help establish participants' mental models. The primary, directed search task was: "Find Bill Clinton's mother's maiden name" (the secondary, general purpose browsing task was: "Find all the information on the relationship between carbon monoxide and desertification"). Following the completion of the task/s, participants had their onscreen search actions played back to them and they were asked why they had engaged in these actions (focusing on the WSE actions). Notes of their responses were taken by the researchers during this process. All participants started the task from the same webpage. After each participant had completed the task the researcher cleared all browsing data (browsing history, download history, emptying the cache, etc) to make sure the following participant could not follow the search path. Participants were then asked to provide an illustration of how they thought a WSE worked accompanied by a brief written description to complement their illustrations. Active participation in this process (from the researcher briefing participants through to the drawing of how a WSE works) ranged from 15 to 55 minutes.

2.3 Measures and Analysis

As this research study was comparing performance for the two samples from 2000 (T1) to 2010 (T2) the number of steps and time taken to complete the directed search task were analysed using a t-test [14]. A chi-squared test was used to assess the accuracy of the answer given by participants comparing T1 to T2 [14]. The salient features were identified from a combination of content analysis of the retrospective verbal protocols and participant drawings/descriptions, Marchionini's [15] conceptual framework of the electronic information-seeking process (i.e. choose search system → formulate query → execute search → examine results), the emergent features from the 2000 data [8], and empirical studies on WSE functioning (e.g. [5][16][17][18]). Two raters working independently assigned the 17 salient features to participants (the 17 salient features are provided in the results). The weighted Kappa coefficient of 92% (91% in 2000) indicated a high level of agreement between the two raters. All remaining discrepancies in the allocation of salient features was done by consensus. In order to determine the mental model clusters for the 2010 sample Wards cluster analysis was conducted using the same salient features identified in the original study as no new features had emerged. As the clusters that emerged were not identical at T1 and T2 a descriptive analysis using chi-squared comparisons of the individual salient features of the respective samples (i.e. potential differences in salient feature 1 at T1 vs. T2) was used.

3 Results

3.1 Performance Measures

Significant differences were found for all the performance measures, with the T2 sample requiring significantly fewer steps and less time to complete the directed search task. Since the T2 sample was significantly more experienced using the Web, experience was added as a covariate. The performance effects were still significant after accounting for experience, suggesting that the performance improvements were due to some other factor such as improved mental models or improvements in WSE functioning rather than increased experience directly. The T2 sample was also significantly more accurate in the answers they provided.

Table 1. Performances differences on the search task from T1 to T2

	Mean T1 (SD)	Mean T2 (SD)	t-statistic
Time since using Web (months)	36.86	119.42	14.82**
Weekly usage of Web (hours)	7.35	16.44	5.49**
Self-rated Web exp. (1-5 scale)	3.32	3.70	2.86*
Steps	14.25 (7.80)	8.48 (4.97)	5.59**
Time	497.10 (272.6)	222.20 (155.9)	7.83**
Correct answers	N=46	N=59	4.68*

** $p < 0.01$; * $p < 0.05$.

The WSE landscape was vastely different from T1 to T2. Participants at T2 either used Google (N=69), Bing (N=8), or Yahoo (N=1) as their WSE to complete the task. At T1 participants used a much wider range of WSEs including Yahoo (N=29), Altavista (N=9), Looksmart (N=9), Infoseek (N=9), Askjeeves (N=8), Google (N=3), Lycos (N=2), Hotbot (N=1), Excite (N=1), Dogpile (N=1), and Metacrawler (N=1). Of course, some of these WSEs were no longer in operation at T2.

3.2 Cluster Analyses of Salient Features (T1 vs. T2)

The 3 clusters for the T2 sample shared a few similarities with the 4 clusters found in the T1 sample. Like in the T1 clusters, a clear indication of an increment in the quality of mental models across the T2 clusters was present. The clusters (1 and 2 at T1, N=17 and N=16 respectively and 1 at T2, N=27) with the simplest mental models in each respective sample showed little understanding of most of the salient features, with a poor understanding of a WSE as a database and the existence of multiple WSEs. However, unlike clusters 1 and 2 in the T1 sample, where there were no participants who showed an understanding that WSE results were ranked, close to 50% of cluster 1 participants in the T2 sample demonstrated this understanding. That is, even in the cluster with the simplest mental models at T2, the T2 sample had slightly more complex mental models. Indeed, there were fewer participants in this cluster at T2 (N=27) compared to the two simplest mental model clusters at T1 (N=33). Cluster 2 of the T2 sample (N=34), which demonstrated more advanced mental models than cluster 1, indicated a substantial improvement regarding the grasp of salient features. Cluster 2 of the T2 sample represents a greater appreciation for Boolean logic and a greater understanding of search terms matching a webpage or database over and above the qualities shared with cluster 1 of the T2 sample. Cluster 3 of the T1 sample (N=21) emphasised the fact that different WSEs produced different results. Cluster 3 at T1 was thus analogous to cluster 2 at T2.

Regarding the most advanced clusters for both samples, a clear appreciation of WSEs as databases that collect information from the Web was apparent. Also, the importance of keywords and knowledge that WSEs rank results was present in both samples. Cluster 4 of the T1 sample (N=26) emphasised the importance of keywords matching with databases (as did a large proprortion of the cluster 2 T2 sample). These qualities were also emphasised in cluster 3 of the T2 sample (N=18), along with more advanced features such as the importance of changing/modifying search terms to influence the quality of the search outcome and the recognition of multiple WSEs (although this latter feature was not as strongly present as for clusters 3 and 4 in the T1 sample).

3.3 Presence of Each Salient Feature: T1 to T2

Salient features more prevalent at T2. Significantly more participants at T2 chose a WSE instead of allowing the Web browser to decide. At T1 a large proportion of the respondents did not have a WSE preference and instead clicked on the browser's "search" button. At T1 this had the effect of the browser randomly assigning one of four different WSEs to the user. Significantly more participants at T2 showed an understanding of the Web as the collection for a database. The participants recognised that the WSE did not search the entire Web (perhaps because the Web is known to be practically too large to search in its entirety and because it changes so rapidly) but gave a representative sample of the Web. More participants at T2 chose key words or phrases to perform the search task. This is because at T1 more participants chose WSE categories and did not enter search terms at all (effectively allowing the WSE to refine the search using predefined categorisation algorithms). At T2 the emphasis was more on user-defined search terms. Similarly, a greater number of participants at T2 showed an appreciation of the impact reformulating search terms had on modifying (widening or narrowing) the search. A larger number of participants at T2 indicated an understanding that WSEs matched search terms and phrases to indicators/tags that the WSE had defined. The use of bold and highlighted terms in WSE summary results probably explains why this occurs. More participants at T2 showed an understanding that WSE algorithms rank results. This was evident in the fact that very few participants went beyond the first page of results and often did not scan below the first four or five search results. More participants from T2 showed an understanding that different WSEs have different ranking algorithms even if this number was, technically speaking, quite small (N=7 at T2 and N=0 at T1). A small number of respondents recognised the differences between WSEs and were more likely to use different WSEs to verify results.

Salient features more prevalent at T1. More participants at T1 understood that different WSEs have different database algorithms (i.e. use different WSE algorithms to collect records from the Web to form a database). Participants at T1 were also more likely to browse the WSE categories to widen/narrow the search domain before submitting search queries (and in some instances did not even submit search query terms). WSEs at T2 largely hide the WSE categories, placing them near the top of the screen and away from the main searching area. Participants at T1 were more likely to use Boolean logic and other operators to change their search parameters. This is primarily because the most common Boolean logic term ("AND" was automatically built into most WSEs at T2). T1 participants were more likely to understand that WSEs match terms and phrases to actual words or phrases in webpages themselves (i.e. not to words and phrases that WSEs have catalogued and tagged). More participants at T1 understood that WSEs match terms and phrases with the webpage title. Participants at T2 appeared to focus more on the search terms being prevalent in the actual document content rather than just in the headings (or meta-tags). Significantly more participants from T1indicated an understanding that WSEs allow users to narrow search results by finding "similar" results (other relevant/connected results). This might be because WSEs have improved their database collection methods to the point where results are found without having to resort to "similar" search terms. Some WSEs also have a scroll-down bar with search term hints. It could be that these functionalities replace the need for searching for "similar" results. More participants from T1 understood that WSEs display a hyperlink to the original location of the information. The number of participants who knew this at T1 was still relatively small (N=22), but significantly larger than at T2 (N=1). This could be because highlighting/bolding the search terms captures greater visual attention and less attention is directed towards the hyperlink.

Salient features the same at T1 and T2. Three salient features remained statistically unchanged. Participants were equally unlikely to recognise that a WSE actually searches a collection from the Web (rather than the whole Web). Likewise, respondents were equally unlikely to recognise that a WSE would use a different collection algorithm to construct the database from the Web (although in both these instances, respondents at T2 reflected slightly more instances of this salient feature than at T1). Finally, no respondents at T1 or T2 understood that a WSE also searches for alternative combinations and extensions of search terms. This is a surprising finding given that the dominant WSE often suggests alternative spellings to search terms with the phrase "Do you mean … ?".

Table 2. Salient features T1 vs. T2

Salient feature	Frequency T1	Frequency T2	χ^2
Chooses a WSE rather than letting the web browser decide (SF1)	36	74	42.00**
Recognises that a WSE is a database (SF2)	10	17	N.S.
Understanding of the web as the collection place for a database (SF3)	26	61	30.86**
Understands that WSEs use a specific algorithm to collect the database (SF4)	8	12	N.S.
Recognises that different WSEs have different databases (SF5)	73	9	102.46**
Participants choose keywords or phrases to search (SF6)	22	73	67.39**
Participants browse WSE categories to widen/narrow the search domain (SF7)	11	3	5.00*
Participants use Boolean logic or other operators to change search parameters (SF8)	52	16	33.14**
Understanding that search terms can be changed/modified to widen/narrow the search of the Web or database (SF9)	6	39	33.66**
WSE matches terms/phrases to the identifiers on the webpage or database (SF10)	4	41	42.32**
WSE matches terms/phrases to words in webpage or Web document (SF11)	54	2	74.28**
WSE matches terms/phrases with the webpage title (SF12)	12	1	10.13**
WSE looks for terms/phrases and related terms/phrases or extensions (SF13)	0	0	N.S.
WSE algorithm ranks results (SF14)	4	30	25.24**
Different WSEs have different ranking algorithms (SF15)	0	7	7.32**
WSEs allow users to narrow search results by finding "similar" results) (SF16)	18	1	17.26**
WSE displays a hyperlink to the original location of the information (SF17)	22	1	22.39**

** $p < 0.01$.

4 Discussion

While seven of the seventeen salient features were more frequent at T1 and the same number were more frequent at T2 there have been clear improvements regarding the completeness and accuracy of WSEs. First, the marked improvements in performance by the 2010 sample provide indirect evidence that mental models of WSEs have improved over the last decade. These performance improvements give an indication that users' interactions with the system have truly become more efficient partly as a function of greater exposure and frequent use; a finding consistent with that of Muramatsu and Pratt [7]. Regarding the equal distribution of salient features, there has been a clear, significant improvement in the comprehension of the various salient features that make up a WSE a decade later. Thatcher and Greyling [8] identified four clusters, two of which displayed highly incomplete and inaccurate mental models of WSEs, only the first cluster of the 2010 sample shared that quality, although even in that cluster, mental models were slightly more complete. The remaining two clusters (52 participants in total at T2 compared to 47 participants in clusters 3 and 4 at T1) showed more accurate and complete mental models of WSEs. The chi-squared comparisons are an indication of the state of WSEs in their respective eras and the demands they placed on their users. A large proportion of users in more recent times have gravitated towards Google with considerable success [19]. Thatcher and Greyling [8] commented on how WSEs did not make many of their salient features transparent and thus made it difficult for users to develop accurate and complete mental models. It would seem not much has changed in that regard as the number of salient features the respective samples were able to identify remained the same. This concern is minor and distracts from a far more important point. It would seem that even though WSEs still hide a great number of their features, they have configured their design specifications to such a high level that this need has been largely neutralised. It would seem that users have become faster, more efficient, and more accurate in their search processes despite still holding relatively naive mental models. It would appear that the salient features that occurred less frequently at T2 are either primarily technical or they are simply unimportant for search performance.

The poor user query formulations noted by a number of authors (e.g. [2][9][12][13]) from which they inferred that inaccurate information was consistently being retrieved was not demonstrated as a concern because of the simpler, if still fairly ambiguous, nature of WSE interfaces. Previous research has been critical of WSE interfaces, suggesting that they fail to deal with cultural and situational contexts as well as the naturally iterative (i.e. users changing, refining, or expanding their search criteria) nature of the search process [19]. Rose [20] suggests that WSE interfaces should be redesigned to invite the refinement of search terms, the use of exploratory searching, and should allow users to select search contexts depending on their search goals. Vaughan and Resnick [21] propose that WSE interfaces should also allow users to easily view their search history as this helps them understand what new search queries to try. As was reported in the introduction, Muramatsu and Pratt [7] found that making the WSEs' transformation of users' queries more transparent assisted in more accurate mental models and improved search performance. WSE designers have already made several modifications (e.g. crawlers driven by 'relevance' principles, drop-down menus, larger databases, automatic search term/phrase correction feedback, etc) that have meant users get the most out of searching without fully comprehending much about the underlying mechanics of WSEs. The mental models held by users in the T2 sample, whilst reasonably better than those of the T1 sample, were still not accurate and complete. It is likely that the decreased time taken to find answers and the reduced number of steps is related to these WSE interface modifications. However, increased download speeds and the increased amount of information on the Internet would also have also contributed in this regard. We would argue, that despite these improvements there are still opportunities for WSE interface designs to become more transparent, to allow more accurate and complete mental models and thereby improve search performance further. There are a number of implications both practically and theoretically from these results.

First, it would seem that having a better mental model of a particular system (even if only moderately better) does in fact help in improving performance in using that system. Second, the results also suggest that systems can be designed to negate the need for highly accurate and complete mental models. However, considering that systems can seemingly overcome the need for effective user mental models for optimised system use there is a theoretical question concerning the real importance of mental models. A large body of literature spanning multiple disciplines has consistently proposed that accurate and complete mental models are essential for effective system engagement [2][22][23]. Indeed, the mediocre search performance of the T1 sample would also give credence to that particular argument. However, the marked differences in performance, even in the presence of the relatively modest improvements in mental models of WSEs, indicate that this is not necessarily true. So, it is possible that high level mental model formation is not always necessary for effective system use, at least regarding directed search tasks. Third, the results show that users' mental models have increasingly aligned better with designers' conceptual models, albeit marginally.

While every attempt has been made to match the T1 and T2 samples the best that can be achieved with this research design is a contrast group. It would have been ideal, from a longitudinal research design perspective, to have sampled the same 80 people ten years later. However, given that participants were anonymous at T1, such a longitudinal design was not possible. It is possible that any variations in performance and mental models were due to using different participants at T1 and T2. The sample size, while relatively good for this type of laboratory-based investigation, is still relatively small in comparison to the population of WSE users. The generalisation of the study findings to other users therefore requires further verification. This study used only one type of search task to determine search performance (i.e. a directed search task determined by the researchers). It is likely that search behaviour varies according to the type of search task and whether it is internally or externally assigned. Further research would have to be conducted on the performance of Web searching using a greater variety of search tasks. Finally, it is worth noting that it is difficult to causally link Web search performance to mental model complexity and accuracy in this study. Search performance may just as easily have improved due to drastically increased download speeds, the enormous increase of available Web content, and improvements in the algorithms that WSEs use to collect, categorise, and rank their databases. Despite these limitations, the study clearly demonstrates that users' mental models of WSEs are now more complex, complete, and accurate.

References

1. Jansen, B.J., Spink, A.: How Are We Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs. Inf. Proc. Manag. 42, 248–263 (2006)
2. Slone, D.J.: The Influence of Mental Models and Goals on Search Patterns During Web Interaction. Journal of the J. Am. Soc. Inf. Sc. Techn. 53, 1152–1169 (2002)
3. Thatcher, A., Greyling, M.: Mental Models of the Internet. Int. J. Ind. Erg. 22, 299–305 (1998)
4. Makri, S., Blanford, A., Gow, J., Rimmer, J., Warwick, C., Buchanan, G.: A Library or Just Another Information Resource? A Case Study of Users' Mental Models of Traditional and Digital Libraries. J. Am. Soc. Inf. Sc. Techn. 58, 433–435 (2007)
5. Crudge, S.E., Johnson, F.C.: Using the Repertory Grid and Laddering Technique to Determine the User's Evaluative Model of Search Engines. J. Document 63, 259–280 (2007)
6. Efthimiadis, E.N., Hendry, D.G.: Search Engines and How Students Think They Work. In: Proceedings of the 28th Annual international ACM SIGIR Conf. Res. Dev. in Inf. Ret. (2005)
7. Muramatsu, J., Pratt, W.: Transparent Queries: Investigating Users Mental Models of Search Engines. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 217–224 (2001)
8. Thatcher, A., Greyling, M.: Mental Models of Search Engines: How Do Search Engines Work? In: Harris, D., Duffy, V., Smith, M., Stephanidis, C. (eds.) Human-centred Computing: Cognitive, social and ergonomics aspects. Lawrence Erlbaum Associates Inc., Mahaw (2003)
9. Zhang, Y.: Undergraduate Students Mental Models of the Web as an Information Retrieval System. J. Am. Soc. Inf. Sc. Techn. 59, 2087–2098 (2008)
10. Liaw, S., Huang, H.: Information Retrieval From the World Wide Web: A User-Focused Approach Based on Individual Experience With Search Engines. Comp. Hum. Beh. 22, 501–517 (2006)
11. Spink, A., Jansen, B.J., Blakely, C., Koshman, S.: A Study of Results Overlap and Uniqueness Among Major Web Search Engines. Inf. Proc. Manag. 42, 1379–1391 (2006)
12. Spink, A., Jansen, B.J., Ozmultu, H.C.: Use of Query Reformulation and Relevance Feedback by Excite Users. Internet Res.: Elect. Net. App. Pol. 10, 317–328 (2000)
13. Spink, A., Wolfram, D., Jansen, B.J., Saracevic, T.: Searching the Web: The Public and Their Queries. J. Am. Soc. Inf. Sc. Techn. 52, 226–234 (2001)
14. Huck, S.W.: Reading Statistics and Research, 4th edn. Pearson, Boston (2004)
15. Marchionini, G.: Information Seeking in Electronic Environments. Cambridge University Press, New York (1995)
16. Bar-Ilan, J., Keenoy, K., Yaari, E., Levene, M.: User Rankings of Search Engine Rankings. J. Am. Soc. Inf. Sc. Techn. 58, 1254–1266 (2007)
17. Crystal, A., Greenberg, J.: Relevance Criteria Identified by Health Information Users During Web Searches. J. Am. Soc. Inf. Sc. Techn. 57, 1368–1382 (2006)

18. Sullivan, D.: How Search Engines Work (2007), `http://searchenginewatch.com/2168031`
19. Search Engine Watch. Top Search Providers for September 2010 (2010), `http://searchenginewatch.com/3634991`
20. Rose, D.E.: Reconciling Information-Seeking Behavior With Search User Interfaces for the Web. J. Am. Soc. Inf. Sc. Techn. 57, 797–799 (2006)
21. Vaughan, M.W., Resnick, M.L.: Search User Interfaces: Best Practices and Future Visions. J. Am. Soc. Inf. Techn. 57, 777–780 (2006)
22. Doyle, J.K., Ford, D.N.: Mental Models Concepts Revisited: some Clarifications and a Reply to Lane. Syst. Dyn. Rev. 15, 411–415 (1999)
23. Senge, P.M.: The Fifth Discipline. Currency Doubleday, New York (1990)

The e-Progression in SEs

Karl W. Sandberg[1], Olof Wahlberg[2], and Fredrik Håkansson[1]

[1] Mid Sweden University, Institution of Information Technology and Media,
851 70 Sundsvall, Sweden
[2] Mid Sweden University, Institution of Social Sciences,
851 70 Sundsvall, Sweden
{karl.w.sandberg,olof.wahlberg,fredrik.hakansson}@miun.se

Abstract. The development of Information and communication technology (ICT) has changed the action of business. The view to considered SEs sector as homogeneous, within which SEs take an ordered, sequential e-progression on the route to ICT adoption, and postulate that businesses move in stages from basic use of the Internet to the full integration of business systems and redesign of business processes. The aim of this paper is to conduct an analysis of the stage model in the context of the progression of ICT adoption by SEs. Empirical cases are given that show weaknesses of stage models to explain e-progression in SEs, the stage model are too general and do not take into account the diversity of SEs and focused upon factors such as firm size, age, owner/manager characteristics and geographical position This variety of different perspectives on the adoption of ICT by SEs suggests the need for a multidimensional framework to more adequate explained e-progression in SEs.

Keywords: SEs; e-progression, stage model, ICT adoption.

1 Introduction

The rapid development and spread of the Internet and related technologies has created new opportunities for SEs. This has placed considerable pressures on SEs in order to adopt the ICT. It has been widely acknowledged that SEs contributes substantially to national economies in both sustaining and creating employment, income and prosperity [1] [2]. In part this is likely to be a result of a lack of awareness of the skills needed and the importance that these skills play in developing a successful e-progression strategy and implementation [3] [4] [5].

Rogers [6] argues that successful adoption of ICT innovations requires a level of knowledge is not just awareness but an actual clear perception of the value ICT can offer that is needed. SEs is particularly constrained by resource factors, and is therefore more sensitive than larger organizations to adoption costs [7] [8]. If there are no clear benefits in ICT adoption, SEs will be more constrained in adoption than a larger company. Many SEs remain in the early stages of Internet adoption, using this technology for lower level functions, not progressing to full integration of the Internet with business processes and lacking clear strategies to identify the potential business

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 254–262, 2011.

value connected to e-business [9] [4] [10]. SEs may avoid adoption to minimising their investment in an unknown innovation [11] [12]. The slow progress of ICT adoption in SEs often focused to owner-manager's characteristics [13].

In developing understanding of how ICTs are developed in SEs, broad research questions framing the data collection and analysis are: How does the owner-manager encourage ICT adoption; how do internal factors in SEs affect ICT adoption; or what are the roles of the owner and other key workers in this process?

The wide spread use of stage models give rise to several questions. Are the categories of a stage model well chosen? Does e-progression in SEs evolve through such a series of stages? Is there a real advancement between the different stages? Should one always strive for higher stages? Are higher stages inherently better than lower stages?

2 Theoretical Framework

Reflection on previous research on SEs e-progression, and also research on ICT adoption, shows that the factors driving adoption of ICT among SEs, owner/managers readiness, external pressure and perceived benefits seem to be the most often applied variables [14]. Several studies have explored the fact that many SEs owner/managers seem to lack knowledge about the potential benefits of ICT adoption and that this lack of knowledge keeps them from getting involved [15]. Studies have suggested that SEs owner/managers, pragmatic and cost- cautious as they are, tend to adopt ICT in smaller steps [16] [17] [18].

2.1 Stage Model for e-progression in SEs

The concept of stage models was originally founded to discuss the adoption and maturity of information systems management strategies from the operational level to the strategically level of organisations [19].

The stage model divided the ICT adoption into several stages; all of them bear a characteristic in describing a development from simple information to e-business integration. Table 1 provides one view of the stages that would be sequentially passed through. The model suggests that SEs have to acquire a threshold level of knowledge relating to e-progression before being able to move onto the next stage of development [20] [21].

These stage models imply that businesses move in stages from the basic use of the Internet, to more sophisticated usage where greater business value will be accrued. There have also been articles that attempt to identify and describe the different phases that SEs moves through with respect to the sophistication of their use of ICT [23]. For example, in relation to website adoption within the SEs context, [24] proposed a stage model relating to the websites they adopt: 1) Static websites, 2) Interactive websites, 3) Interactive/transactional websites, and 4) Fully integrated websites.

Table 1. Stages of e-progression in SEs [22]

Stage	ICT-adoption	Description
0	*Not started*	The business does not have Internet access.
1	*E-Mail*	Accesses information and services on-line and uses e-mail. Does not have website, or surfs the Web, but has an efficient internal and external communications structure.
2	*Website*	Business has website but contains only basic information about business, relies on customer initialising contact for further information. Can buy services and supplies on-line.
3	*E-commerce*	Customers have access to more information (catalogue) about products/services. On-line ordering and payment (store) system. Reduced costs and higher levels of accessibility and speed. Website not linked to internal systems and orders are processed manually.
4	*E-business*	Have integrated supply chain, ordering, manufacture, delivery, accounts and marketing to other business systems (seamless processing). Minimum (reduction of) waste regarding resources between supply chain stages.
5	*Transformed organizations*	Open information systems for customers, suppliers and partners. Internet technology drives both external and internal processes more effectively and efficiently (enabled). Based on networking between firm and other organizations/individuals.

2.2 Disadvantages of Using Stage Model Adoption of ICT in SEs

A main contribution of the paper is to propose an alternate to the stage model to explain e-progression in SEs. We will achieve this through a combination of conceptual and empirical analyses. Encountering such shortages when attempting to make the move to the next stage of development may delay and constrain the business to some extent. It would be expected that performance of the business will be most adversely affected if the business makes the leap to the next stage, but finds itself without the skills required to operate at this level. In this manner the severity of the consequences of any skill shortage will be expected to lag the stage of development when firms are most likely to encounter the skills shortage.

Significantly, however, no further flexibility is built into the model in order to encompass the impact of key factors such as size, sector, ethnicity, gender, human and financial resources, customer base, adoption stage and level of internationalisation. Typically, e-commerce was found to benefit innovative SEs that initiate and develop new types of business relationships [25] [26].

Martin and Matlay [22] also offer a critical analysis of previous research as well as the implicit assumption in the adoption approach that all SEs somehow can subscribe to a stage development in ICT. The stage model of ICT adoption by SEs may itself be problematic. Most attempts to use the model to explaining SEs economic activity tended to oversimplify complex issues and circumstances [27] and their effectiveness and generalization has also been questioned [28]. In their view, this generalist view of

SEs operation fails to distinguish between businesses of various sizes, owner/manager characteristics, stages of adoption and so on. Blackburn et al [13] found that owner-managers were a key influence in determining use, based on attitudes to ICT, level of ICT skills, and management orientation. Mendo et al [29] identified weaknesses of stage model to describe the adoption of ICT by SEs: An oversimplified perspective of complex issues and circumstances, based on a false assumption that SEs progress from basic to more advanced adoption of ICTs in a linear fashion, a lack of empirical validation, a generalization that does not take into account the diversity of small businesses, and a focus on the broad picture of change in the SEs, rather than individual instances.

A number of researchers have attempted to develop contingent role models as alternatives to staged models of ICT adoption. These models are based on the premise that different types of business will view ICT adoption in different lights. Tagliavini et al [30] identified following ICT adoption for e-commerce in following way: 1) Public relations, 2) Company promotion, 3) Pre/post sales support, 4) Order processing, 5) Payment management.

This variety of different perspectives on the adoption of ICT by SEs suggests the need for a broader multidimensional framework to be adopted [29]. The key point here is that this content of change context does not fall into the stage model trap, of presuming that the ICT SEs adopt will go through a phased staged process in a linear fashion, does not present that suggests certain levels signify more advanced thinking in adoption - for example not all SEs will want to transact online, which does not mean that they are less developed in their thinking on adoption.

Fillis [31] has questioned stage models and suggested that the adoption of ICT is a non-linear process. It may fail to illustrate the process that may take place at micro level within individual SEs [32]. Effective adoption and implementation of ICTs in SEs may rely more on individual factors such as owner/manager characteristics, structure and mix of available resources. Training for skill ownership at particular ICT levels are implied in the background to the stage model but not identified explicitly as rungs on the stage-to-stage progression to implementation in SEs.

Other difficulties in using stage model, generalised models might result from its application to a target that exhibits the inherent diversity of a small business sector, where variety can lie in size, economic activity, geographic position, resource availability. Penrose [33] and Hawkins et al [34] pointed out that because of its size and diversity, research results cannot be easily generalised across all SEs.

The relatively smaller size of the average SEs organisation has been identified as the main factor for lower adoption rates of ICT. It has been established that SEs are least likely to be involved with e-commerce and/or the Internet. Those SEs that specialise in the provision of business services, particularly knowledge intensive service organisation are more likely to adopt ICT than similar sized manufacturing SEs [35].

Given the emphasis on wider and more rapid adoption of ICT in countries with dispersed rural populations [36] it follows that the experience of companies in rural and urban areas might also differ. Rural communities are expected to benefit considerably from the recent expansion of "commercial niche market" opportunities, including a growing interest in locally sourced food, marketed mainly via the Internet [36] [37]. Only in the few cases where rural extranets had been set up did SEs mimic the stage model, using communication by e-mail as their first implementation step and moving on to selling to each other as part of this process [38]. The difficulties faced by rural firms are not reflected in the stage model, nor are their experiences identified as a potentially different or alternative process.

The reactive or proactive approach of owner/managers to rapid technological changes in the marketplace is crucial to ICT adoption and implementation and perceptions of ICT benefits are key features in this process [39]. In those SEs that lack international awareness, the ideas, awareness and ICT capacity of managers and decision makers may also need to be expanded. Thus, those SEs with lower ICT understanding and knowledge usually have considerable problems in identifying and/or fully appreciating the usefulness and easy to use of ICT at implementation and at each stage of development.

In studying ICT adoption, the owner-manager is an important source of information on the process. Owner-managers are usually cited as the central point for SEs; their role in innovation and change relates to the inter-relationship between their own attributes and the characteristics of the SEs and to their dominance in decision making [40] [11]. The problem may be that SEs may not operate in logical ordered and sequential ways, referred to stage models. To take a more overlapping approach with a learning and knowledge focus, alternative models may also have a greater synergy with the reality of e-progression in SEs [41] [42].

3 Survey of SEs Adoption of ICT

In this study, we define e-progression in SEs as the sharing of business information, maintaining business relationships, and conducting business transactions by means of ICT. Although sometimes used interchangeably, e-commerce is traditionally defined as a subset of e-business which concerns only business transactions accomplished using internet-based technologies [43]. By applying a broader definition of the concept of e-progression, we also included customer service provided by developing and maintaining ICT. The study is a part of a going on project “The Digital Age in Rural and Remote Areas” DARRA [44]. In total, 35 rural manufacturing and services SEs in Sweden participate in the study. The owner/managers of the SEs are asked whether they use of e-mail, web-site, e-commerce, e-business and transformed organisations.

3.1 Results from Survey

Result in figure 1 shows that all SEs use e-mail, nearly all (94 %) of SEs use website, 39 % of the SEs use e-commerce, e-business staying for one-fifth (20 %) of SEs, and nearly one-third (29 %) of SEs networking with other business partner.

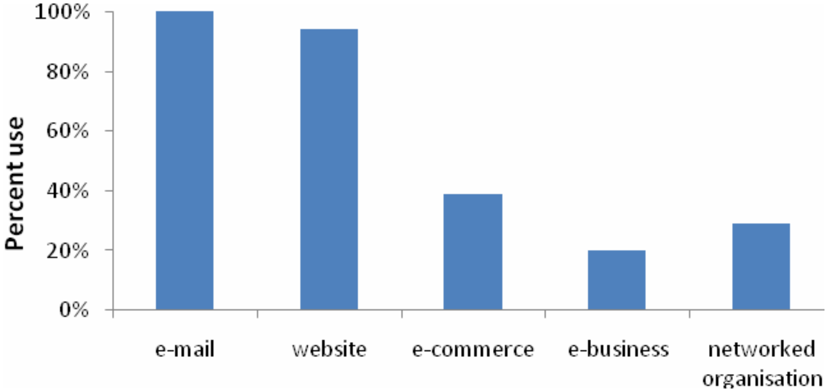


Fig. 1. Percent use of the ICT in SEs

Based on previous research, in order to deal with the general uncertainty concerning potential returns on ICT investments, SEs owner/managers tend to adopt e-commerce in smaller steps [23] [17]. The survey confirms this behaviour. Fewer SEs use ICT in e-commerce.

4 Discussion and Conclusions

We have in this paper investigated shortcomings of stage models for e-progression in SEs. The most advanced stage in stage model does not always include all lower stages. The main benefits of e-business are not in the early stages of adoption but rather in the more sophisticated applications, such as online business processes.

The difficulty that the SEs faces is that they are generally too small to have an in-house team dedicated to updating and maintaining ICT. This makes it unlikely they will be able to find the skills required to move towards becoming a true e-business, however, whilst SEs who had limited Internet use within the firm were most likely to suffer from basic IT skills shortages.

Pippen [45] and Darch et al [46] suggest that shortages of IT skills in general and those specifically related to e-commerce are widespread appear to be confirmed by the paper's findings.

The stage model adoption of ICT in SEs outlined in this paper represents both a limited and a limiting vision of government-inspired support for the ICT implementation and development needs of firms operating in the SEs sector of the Swedish economy. The stage model needs to be revised, extended and modified so that it reflects more fully the requirements and experiences of firms of various sizes and geographical locations, and in particular of those organisations that exhibit more entrepreneurial or sophisticated usage of ICT. SEs will struggle if they attempt to integrate ICT into their firm without owner/managers and their workforce having relevant knowledge and adequate human resource capacity to support ICT. The importance of human capital for ICT acquisition and development should be explicitly recognised in this model and it may also prove useful to replace the stages of adoption with more suitable stages of SEs understanding. This is particularly important during initial stages, as the commitment of owner/managers and their perception of ICT benefits appear to be crucial to the successful adoption and development of new ICT technology in this type of firm.

For those struggling to adopt and in the earliest stages of ICT development it would perhaps be most appropriate that training schemes be made widely available for existing SE employees, as few SEs will be in the position to take on new dedicated staff at such embryonic stages. Although there is a fear that this investment is not tied to the business, there may also be benefits including lower staff turnover and higher job satisfaction [47].

SEs were identified as possible targets for support as they are most likely to benefit from ICT, through their focus on "niche markets" and their tendency to develop and maintain close relationships with both customers and suppliers. In this context, new ICT resources could provide SEs with key competitive advantage and facilitate the search for underdeveloped and intrinsically lucrative outlets in both the domestic and the global market.

For the finishing conclusion of ICT-related initiatives it is crucially important to recognise of specific needs, strategies, ideas and core capabilities that coexist in this sizeable sector of the economy. There is an acute need for further quantitative and qualitative research on this important topic. More empirical research is needed at micro-economic level to facilitate a better understanding of the complex processes and differentiating factors that affect ICT adoption levels and its impact upon SEs competitiveness. Without better understanding, the drive for ICT adoption and development will not successfully contribute to e-progression in SEs in the future.

References

1. Beaver, G.: Small Business. Entrepreneurship and Enterprise Development. Pearson Education Limited, Harlow (2002)
2. Jutla, D., Bodorik, P., Dhaliwal, J.: Supporting the e-Business Readiness of Small and Medium - Sized Enterprises: Approaches and Metrics. Internet Research; Electronic Networking Applications and Policy 12(2), 139–164 (2002)
3. Lin, C., Huang, Y.-A., Tseng, S.-W.: A Study of Planning and Implementation Stages in Electronic Commerce Adoption and Evaluation: The Case of Australian SMEs. Contemporary Management Research 3(1), 83–100 (2007)
4. Ramsay, E., Ibbotson, P., Bell, J., Gray, B.: E Opportunities of Service Sector SMEs: An Irish Cross Border Study. Journal of Small Business and Enterprise Development 10(1), 250–264 (2003)
5. Beckinsale, M., Ram, M.: Delivering ICT to Ethnic Minority Businesses: An Action-Research Approach. Environment and Planning C: Government and Policy 24(6), 847–867 (2006)
6. Rogers, E.M.: The Diffusion of Innovations, 4th edn. Free Press, New York (1995)
7. Lewis, R., Cockrill, A.: Going Global - Remaining Local: The Impact of e-Commerce on Small Retail Firms in Wales. International Journal of Information Management 22(15), 195–209 (2002)
8. Van Smith, J., Webster, L.: The Knowledge Economy and SMEs: A Survey of Skills Requirements. Business Information Review 17(3), 138–146 (2000)

9. Southern, A., Tilley, F.: Small Firms and Information and Communication Technologies (ICTs): Toward a Typology of ICTs Usage. New Technology, Work and Employment 15(2), 138–154 (2000)
10. Windrum, P., de Berranger, P.: The Adoption of Intranets and Extranets by SMEs. Merit Research Memoranda 2003–2023. MERIT, University of Maastricht (2003)
11. Sandberg, K.W., Öhman, G.: The Science Inside Innovation Process Factors. In: Sandberg, K.W., Öhman, G. (eds.) The 3rd Symposium on the Entrepreneurship-Innovation-Marketing Interface and 2nd BIEM-Symposium, June 11-12, Cottbus, Germany (2009)
12. Lange, T., Ottens, M., Taylor, A.: SMEs and Barriers to Skills Development: A Scottish Perspective. Journal of European Industrial Training 24(1), 5–11 (2000)
13. Blackburn, R., McClure, R.: The Use of Information and Communication Technologies (ICTs) in Small Business Service Sector Firms. Small Business Research Centre, Kingston Business School, London (1998)
14. Jeyaraj, A., Rottman, J., Lacity, M.J.: A Review of the Predictors, Linkages and Biases in IT Innovation Adoption Research. Journal of Information Technology 21(1), 1–23 (2006)
15. Dutta, S., Evrard, P.: Information Technology and Organisation Within European Small Enterprises. European Management Journal 17(3), 239–251 (1999)
16. Daniel, E., Wilson, H.: Adoption Intentions and Benefits Realised: A Study of e-Commerce in UK SMEs. Journal of Small Business and Enterprise Development 9(4), 331–348 (2002)
17. Eriksson, L.T., Hultman, J.: One Digital Leap or a Step-by-Step Approach? A Longitudinal Study of e-Commerce Development Among Swedish SMEs. International Journal of Electronic Business 3(5), 447–460 (2005)
18. Drew, S.: Strategic Uses of e-Commerce by SMEs in the East of England. European Management Journal 21(1), 79–88 (2003)
19. Nolan, R.: Managing the Computer Resource: A Stage Hypothesis. Communications of the ACM 16, 399–405 (1993)
20. Levy, M., Powell, P.: Exploring SME Internet Adoption: Towards a Contingent Model. Electronic Markets 13(2), 173–181 (2003)
21. Lee, J.: Discriminant Analysis of Technology Adoption Behaviour: A Case of Internet Technologies in Small Businesses. Journal of Computer Information Systems 44(4), 57–66 (2004)
22. Martin, L., Matlay, H.: Blanket Approaches to Promoting ICT in Small Firms: Some Lessons from the DTI Ladder Adoption Model in the UK. Internet Research: Electronic Networking Applications, and Policy 11(5), 399–410 (2001)
23. Daniel, E., Wilson, H., Myers, A.: Adoption of e-Commerce by SMEs in the UK: Towards a Stage Model. International Small Business Journal 20(3), 253–270 (2002)
24. Rao, S., Metts, G., Monge, C.M.A.: Electronic commerce development in small and medium sized enterprises: a stage model and its implications. Business Process Management Journal 9(1), 11–32 (2003)
25. Trappey, C.V., Trappey, A.J.C.: Electronic Commerce in Greater China. Industrial Management & Data Systems 101(5), 201–210 (2001)
26. Feher, A., Towell, E.: Business Use of the Internet. Internet Research 7(3), 195–200 (1997)
27. Kai-Uwe Brock, J.: Information and Technology in the Small Firm. In: Carter, S., Jones-Evans, D. (eds.) Enterprise and the Small Business, pp. 384–408. Prentice Hall, Pearson Education, Englewood Cliffs (2000)
28. Matlay, H.: Vocational Education and Training in Britain: A Small Business Perspective. Education + Training 41(1), 6–13 (1999)

29. Mendo, F.A., Fitzgerald, G.: A Multidimensional Framework for SME e-Business Progression. Journal of Enterprise Information Management 18(6), 678–696 (2005)
30. Tagliavini, M., Ravarini, A., Antonelli, A.: An Evaluation Model for Electronic Commerce Activities within SMEs. Information Technology and Management 2(2), 211–230 (2001)
31. Fillis, I., Johansson, U., Wagner, W.: A Qualitative Investigation of Small Firm e-Business Development. Journal of Small Business and Enterprise Development 11(3), 349–361 (2004)
32. Fallon, M., Moran, P.: Information Communication Technology and Manufacturing SMEs. Paper presented to the Small Business Enterprise and Development Conference, University of Manchester, April 10–12 (2000)
33. Penrose, E.T.: The Theory of the Growth of the Firm. Basil Blackwell, Oxford (1959)
34. Hawkins, P., Winter, J., Hunter, J.: Skills for Graduates in the 21st Century. Report Commissioned from Whiteway Research. Association of Graduate Recruiters, Cambridge (1995)
35. Sandberg, K.W., Wahlberg, O., Pan, Y.: Owners/Managers Acceptance of ICT Innovation in Small Business. In: Present in Conference ISIT 2009, October 12–13, Novo Mesto, Slovenien (2009)
36. Sandberg, K.W.: Information Technology and Network in Small Enterprises in Rural Area. In: Persson, L.O., Sätre Åhlander, M., Westlund, H.(ed.) Local Responses to Global Changes. Economic and Social Development in Northern Europe's Countryside (2003)
37. Sparkes, A., Thomas, B.: The use of the Internet as a Critical Success Factor for the Marketing of Welsh Agri-Food SMEs in the Twenty-First Century. British Food Journal 103(5), 331–337 (2001)
38. Martin, L.M.: E-Commerce and Existing Small Firms: A West Midlands Pilot Study of SME Internet use. University of Glasgow, Glasgow. Paper presented to the European trade Study group (2000)
39. Poon, S., Swatman, P.M.C.: Small Business Use of the Internet, Findings from Australian Case Studies. International Marketing Review 11(5), 385–402 (1997)
40. Culkin, N., Smith, D.: An Emotional Business: A Guide to Understanding the Motivations of Small Business Decision Takers. Qualitative Market Research- An International Journal 3(3), 145–157 (2000)
41. Dixon, T., Thompson, B., McAllister, P.: The Value of ICT for SMEs in the UK: A Critical Literature Review. Report commissioned by the UK Small Business Service, The College of Estate Management (September 2002)
42. Lawson, R., Alcock, C., Cooper, J., Burgess, L.: Factors Affecting Adoption of e-Business Technologies by SMEs: An Australian Study. In: Lawson, R., Alcock, C., Cooper, J., Burgess, L. (eds.) Present First International Conference on Business Innovation in the Knowledge Economy, June 12, IBM, Warwick (2002)
43. Chaffey, D.: E-Business and E-Commerce Management. Prentice-Hall, Englewood Cliffs (2004)
44. DARRA (2010), `http://www.northernperiphery.eu/en/projects/show/&tid=13`
45. Pippen, A.: Digital Talent for Digital Britain - Will the Next Generation Achieve the Vision? In: Robinson, P. (ed.) eBritian, British Institute of Technology and E-commerce,, pp. 11–13 (2010)
46. Darch, H., Lucas, T.: Training as an e-Commerce Enabler. Journal of Workplace Learning 14(4), 148–155 (2002)
47. Choo, S., Bowley, C.: Using Training and Development to Affect Job Satisfaction Within Franchising. Journal of Small Business and Enterprise Development 14(2), 339–352 (2007)

Cross-Cultural Comparison of Blog Use for Parent-Teacher Communication in Elementary Schools

Qiping Zhang and April Hatcher

Long Island University, Palmer School of Library and Information Science,
Brookville, NY, 11021, USA
qiping.zhang@liu.edu, infogirl84@yahoo.com

Abstract. There are many factors that effect student learning and achievement. Factors such as socioeconomic status, class size, a child's learning style, and parental involvement all have influence on a student's achievement in school. In this study, we focus only on the factor of parental involvement as it relates to parent-teacher communication. Parent-teacher communication has traditionally been conducted through parent-teacher conferences, personal letters to parents, telephone calls home, etc. However, the growth of the Internet based communications such as e-mails and blogs have expanded the ways in which parent-teacher communication can occur. The objective of this study is to find out how blogs, a lightweight web 2.0 technology, are used to support communication between parents and teachers in different national culture settings. The findings of this interview study suggested that cultural values, privacy policies, teacher background and technology knowledge have influenced the use of blog in parent-teacher communication.

Keywords: blog, web 2.0, culture, computer-mediated communication, parent-teacher communication.

1 Introduction

Industrialized nations around the world are striving to educate their populations to compete in today's globalized economy. The public education system within the United States has found itself struggling to keep up with this new educational demand. The U.S. has consistently ranked far behind other nations in academic achievement. Each U.S. state has traditionally been responsible for the education of its citizens. Over the years, however, the federal government has tried to play a big role with the passing of educational initiatives such as the Elementary and Secondary Education Act (ESEA) of 1965 and it's successor, the Bush administration's "No Child Left Behind Act" (NCLB) of 2001. The NCLB act raised testing and learning standards. In 2009, the Obama administration enacted the "Race to the Top" initiative as part of the federal government's continuing efforts to improve student achievement through school reform. These educational reforms require more accountability from within the educational community at local and state levels. However, there are many factors beyond standardized testing that impact upon student learning and achievement. Factors such as socioeconomic status, class size, a child's learning

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 263–272, 2011.

style, and parental involvement all have influence on a student's achievement in school. In this study, however, we focus only on the issue of parental involvement as it relates to parent-teacher communication.

The purpose of this study was to find out how blog is used in k-12 education particularly in supporting communication between parent-teacher communications in different national culture settings. Two major research questions for this study are: what is the role of blogs in parent-teacher communication? What are the cross-cultural differences of blog use in parent-teacher communication?

2 Related Work

In the following we will first review literature on parental involvement, and parent-teacher communication, then on technology aspects of parent-teacher communication including computer-mediated communication in education, and blog usage in classroom. We will end up our literature review with studies on cultural differences in parent-teacher communication.

2.1 Parental Involvement

Researchers have looked at parental involvement with a focus on teachers' parent involvement practices [1, 2, 3]. A 1983 Johns Hopkins study [2] of elementary teachers' parental involvement practices found that teachers' in different types of school districts (urban, suburban, rural) emphasized different types of parent involvement. Overall, the study concluded, "optimal programs for parents result from teachers' frequent involvement of parents in learning activities at home."

There is not always agreement between parents and teachers as to how parental involvement will be carried out. According to Epstein [3], "teachers have strong opinions about parent involvement. Some believe that they can be effective only if they obtain parental assistance on learning activities at home. Others believe that their professional status is in jeopardy if parents are involved in activities that are typically the teachers responsibilities."

Others have looked at defining and measuring parental involvement [4, 5] and it's impact on student achievement [6]. Fan and Chen [6] conducted a meta-analysis of the literature in order to examine the multifaceted nature of parental involvement. They found that "parental aspiration/ expectation for children's education achievement has the strongest relationship, whereas parental home supervision has the weakest relationship with students' academic achievement."

Effective parental involvement relies on frequent communication between classroom teachers and parents. Good teacher-parent communication is frequent, timely, straightforward, and honest [7]. In his study, Powell [8], found that the frequency of teacher-parent communication can depend on parents' attitudes towards teachers and teacher role status. Another aspect of the communication process is differing communication styles among teachers and parents.

While teachers tend to employ institutional communicative methods, parents prefer more personal individual invitations for involvement [9]. Teachers have to find ways to involve parents in their child's education. There are a variety of factors that can impact upon this potential communication. According to Halsey [9], teachers, parents and students are often uncertain about how to initiate parent involvement in their schools. In her study she found that "all participants in the study agreed that parent involvement was beneficial to student success and positive school-family relations, but teachers, parents, and students faced many obstacles when they went about planning and implementing parent involvement." One of the obstacles found was in implementing effective methods of communication. She concludes, "one difficulty in the initiation of parent involvement is that teachers and parents perceive communicative efforts differently." [9]

Kohl, Lengua and McMahon [5] created the Parent-Teacher Involvement Questionnaire (PTIQ) to aid in assessing parent-school partnerships. Seitsinger, et al [10], also developed a Teacher-Parent Contact Scale (TPCS) to measure teacher-parent contact practices to improve parent-teacher contact.

Teacher-parent communication has traditionally been conducted through teacher-parent conferences, personal letters to parents, telephone calls home, etc. The growth of Internet-based communications such as e-mails and blogs have expanded the ways in which parent-teacher communication can occur. Internet-based communication methods offer advantages over more traditional methods in their ability to increase the frequency and outreach of communication between families and schools [11, 12]. In her study, Thompson [13] examined the characteristics of parent-teacher e-mail communication with a focus on the pedagogical consequences of CMC. Another study of the pedagogical implications of CMC was conducted by Kim [14] who proposed a theoretical model of blog use in educational settings. Richardson [15] and Zawilinski [16] discuss the ways blogs are used in elementary and secondary education.

2.2 Parent-Teacher Communication

How teacher-parent communication is carried out is an integral part of the parental involvement process. Traditional forms of parent-teacher communication have included informal methods such as personal notes, face-to-face meetings, personal e-mails, telephone calls and more formal methods such as parent-teacher night, bulletins and flyers to parents, letters to all parents, e-mails to all parents, etc. [7]. The Johns Hopkins study [2] found that most parents indicated they were never involved at school; a large percentage of parents did not receive basic, traditional communications from school to home, such as notes, conversations, phone calls, or conferences with teachers.

According to Epstein [3], "communication from the school to the home is sometimes considered 'parent involvement' but is usually 'parent information." She also points out that, "All schools send information home to the family about schedules, report card grades, special events, and emergency procedures. Most of these activities flow one way from the school to the home, often with no encouragement for communication from parents." [3] A way in which to improve upon parent-teacher communication was the focus of study by Seitsinger, et al [10]. They developed the Teacher-Parent Contact Scale (TPCS) in order to measure teacher contact with parents. In their study, "teacher/school practices in contacting parents were found to be significantly related to parent reports of school contact performance and student adjustment and achievement.” [10]

Banach [7] points out that the message to parents should be interesting no matter what communication method is used. He states that parents are interested in knowing: if their child is safe and secure at school, information about their teacher’s qualifications and what their child is learning. They also want frequent progress reports and they want to know ‘right now’ if their child is struggling academically or ‘in trouble’ at school. Don’t wait until conference time.” [7] He also recommends that teachers take a parent-teacher communication inventory in order to assess their classroom environment and parent community. Teachers should use this inventory while viewing their classroom as a living system because it continually growing, developing, changing and adapting.” As such, “systems in a rapidly changing environment need more communication than systems in a slowly changing environment. Also, “there needs to be more informal communication in rapidly changing environments.”

2.3 Computer-Mediated Communication in Education

Internet-based communications such as e-mail have become a popular medium for parent-teacher communication. For many teachers and parents e-mail offers a more convenient means of contact. Mitchell, Foulger and Wetzel [11] advocate for the use of Internet-based communication methods. They state, “while traditional forms of home-school partnerships (for example, parents participating in class activities and teachers sending home children’s work) are associated with positive results, they are limited in their ability to effectively reach all families.”

Tobolka [12] also advocates for the use of electronic communication because of the problem with paper notices and telephone calls to home. She describes the frustration that many teachers encounter in school to home communication. She states, “ I got tired of writing notes that never made it home, and of students losing school work and homework. I spent hours calling their homes, copying notes and looking for more copies. I needed a new strategy to address communicating from school to home.” The results of her study revealed that with Internet-based technology not only parents felt more involved in their student’s school activities, but also the students responded positively. “Students increasingly felt that it was important that they turned in their work and homework - they all wanted to have positive notes to sent home.”

Similarly Thompson [13] reported, “the use of computer mediated communication such as e-mail has reportedly increased the level of parental involvement and parent-teacher communication at the elementary and secondary level.” As a result of usage of parent-teacher e-mail, homework completion increased and some students achieved higher grades. However, she also found negative consequence of e-mail communication. Given email makes it easier for teachers to report grades to parents, parent-teacher communication becomes even more grade-based rather than discussing the specifics of what students are learning.

2.4 Blog Usage in Classroom

Blogs are used in elementary school settings in different ways. There are four common types of blogs found in elementary classrooms according to Zawilinkski [16]. They are classroom news blogs, mirror blogs, showcase blogs and literature response blogs. The focus of this study is on classroom news blogs which are used to share news and information with parents and students. "Teachers update classroom news blogs on a regular basis, posting homework assignments, providing updates on curriculum for parents, and sharing any other information that could benefit the home-school connection." Richardson [15] states that classroom uses of weblogs allow teachers to "post class related information such as calendars, events, homework assignments, and other pertinent class information. In addition, blogs can be used to "communicate with parents if you are teaching elementary students."

2.5 Cultural Difference in Parent-Teacher Communication

Cultural differences and perceptions can also have impact upon teacher parental involvement practices. Teachers' may have preconceived notions about communicating with diverse families and vice versa [17, 18].

In their study, Huntsinger and Jose [19] compared parental involvement on academic achievement across two specific cultures (European American and Chinese American) within the United States. Their study included three types of parental involvement: communicating, volunteering at school, and learning at home. They found that European American parents volunteered more at school while Chinese American parents focused more on systematic teaching of their children at home.

While classroom size is pertinent as teachers must devise ways to communicate with the parent(s) of each student, the educational system in the United States and Asian countries differ in classroom size dramatically. In the U.S. a smaller teacher to student classroom size ratio is desired at the elementary level even though the teacher-student ratio varies widely from state to state. The teacher to student ratio averages for the 2007-2008 school year was one teacher per approximately twenty students, and that number has been increasing as of late [20]. In Japan and South Korea, the averages are higher at thirty-three and thirty-six respectively [21]. In China, the ratio is even higher with approximately one teacher per every forty-five to sixty students being the average norm. The larger the class, the more parents that need to be contacted. Thus different ways to support parent-teacher communication are needed in different cultures. The goal of the reported study is to answer this question by conducting in-depth interview with elementary school teachers in two cultures that have very different teacher-student/parent ratio: the U.S. and China. We are not only interested in how teachers communicate with parents currently, but also would like to find out how they plan or already have integrated Internet-based technologies especially web 2.0 technologies such as blog into their communication with parents.

3 Method

3.1 Participants

Three elementary school teachers were interviewed for this study: two Chinese females (Ms. Y. & Ms. G.) at the age of 25-40 from the same elementary school in China, and one American male (Mr. B.) from the U.S. at the same age range as above.

Given the cultural differences on educational practice, it is hard to choose two elementary schools with similar school philosophy, pedagogical method, parental involvement etc. Instead to keep the cultural comparability, we chose one elementary school from each culture: China and U.S., that is representative in its own culture in terms of its student population (majority from middle-class families), school attitude to new technologies.

3.2 Survey and Interview Questions

This is a semi-structured interview study. Given the small population of elementary school teachers who use blogs, we decided to conduct in-depth interviews for our study. To keep the consistency in our interview, a survey of blog usage in classroom was developed and used based on previous literature [15] to conduct our interview. In the survey, background questions are to collect demographic and teaching practice information (e.g. grade, education background, technology skills, teaching philosophy); other questions include their blogging practice (motivation, use history, training), teacher-parent communication (channel, benefit, challenge), and privacy issue. The interviews were all conducted over the phone.

4 Results

The interviews were analyzed to address following research questions.

4.1 Blogging Practice

Which blog hosting sites did you use? Why? How long have you been blogging?

Both American teacher and Chinese teachers reported school-hosted or school recommended sites: School World (U.S.) and blog.sina.con (China), that is a popular public blog hosting site in China similar to blogger.com in the U.S.). The reasons for their choices are similar too: easy to use, free, has needed feature. All three teachers reported starting their blog usage in 1.5-2 years ago.

Why did you start your classroom blog? Where did you learn blogging skill?

Similarly, all teachers reported following communication motivations to start their classroom blog, which is to provide an additional channel to communicate with students (expose them web 2.0 applications) and their parents (announce news and send messages to parents).

Differently, Mr. B. reported two additional motivations: to support student to student communication, and being able to have several blogs (for different classrooms) within the site. Chinese teachers reported one additional motivation: to use bog to communication among teaches and between teachers and administrators.

This practice reflects collectivism nature of Chinese culture, in which group interests are emphasized. Even though Chinese started their blogging on their own, not required by the school, they still intended to use it as an additional channel to support their professional communication with their colleagues and their bosses.

In terms of training, all three teachers reported they self taught how to use blog. One Chinese teacher reported that she participated in a regional teacher competition where she heard the concept of blogs. Thereafter she started her own personal blogs at two other popular sites (blog.sohu.com and blog.163.com) before she started her professional classroom blog at blog.sina.com. Mr. B. reported similar experience. He started reading blogs of personal interest and some teacher blogs. Then through his own personal efforts he gained skills of blogging. While blogging was not required to all our three interviewees, their self-imitated blogging behavior reflected their strong self-motivation for professional development, and their braveness to try out new technologies in their teaching practice.

What do you blog?

Mr. B. reported to use blog to announce classroom news, share resources such as supplementary readings, and to extend classroom conversations.

Ms. Y. as a class teacher, showed similar activities to Mr. B. In addition,

- she used blog to show samples of student work. As a matter of fact, it was honor for students to have their work posted on her blog site.
- She regularly took pictures of class activities and posted them to her blog to keep parents informed. In contrast, such posting were forbidden in the U.S. schools because of privacy concern. We will review privacy issues in details in 4.3.
- Interestingly Mr. Y. said that she didn't use her blog to extend classroom conversations due to the concern of efficiency of virtual discussions.
- She blogged differently depends on which grade she was teaching. For 6th grade, (highest grade in Chinese elementary schools), she mainly blogged for students (who knew how to get online and are able to read blog). For 1st grade, where students don't have strong reading nor computer skills to understand blogs, she mainly blogged for parents with school activities.

Ms. G., as a math teacher, is not in charge of daily routine activities in a classroom that is reserved for class teacher in Chinese elementary schools. Therefore she didn't use blog to announce classroom news. Rather she mainly posted materials students don't have time to study in regular class sessions, or advanced / challenging questions for those students who would like to have enrichment experience outside the classroom. Given the average of class size of 45-60, Ms. G. told us that blog is an efficient way for her to empower gifted and advanced students.

How often do you update your blog postings?

Since all three teachers were not required to use blog for their teaching, they reported that in general they updated their blog posting whenever they had free time. However, Chinese teachers said that their students would ask them why they didn't blog if they didn't see new postings in 2-3 days. Therefore they tended to update their blogs within 2-3 days. If they didn't have news or classroom related materials to share, they would search educational websites and forwarded some interesting stories or articles on their blogs to meet students' expectation.

4.2 Parent-Teacher Communication

How do you communicate with parents currently?

Mr. B. had a parent-teacher meeting once per year. He mainly used classroom news to communicate with parents. Given 23 out of 24 family had Internet access, he posted classroom news on his blog site instead of handing out regular papers home. Other than blog, he used telephone and email to communicate with parents.

Ms. Y. reported multiple ways to communication with parents. First, she constantly used her blog to announce classroom news, remind parents of upcoming exams, deadlines so that they could help their kids to prepare for them. Second, she never had one-on-one parent-teacher meetings unless students had troubles. Parent-teacher meetings are once every semester and always held with all parents at the same time. Thirdly, she used text messaging on her cell phone and on Internet (QQ) extensively. Telephone is used for emergency or immediate attention. Finally, she used notebook to notify parents student's academic performance, and regularly required parents' signature upon their inspection of students' work.

What benefits do you gain from blogging compared to other communication channels?

Mr. B. reported benefits including wider audience, less formal, issues not necessarily pressing, allow more feedback. Ms. Y. and Ms. G. reported flexibility (parents can review news and materials afterwards), efficiency, deep conversation like her teaching philosophy.

What challenges do you encounter in using blog compared to other communication channels?

Mr. B. mentioned few challenges he faced, because its use was supplemental. But he feared that once school district catch up with the idea of blog, they might take over and he would lose the freedom of doing what he wanted to blog. Ms. Y. and Ms. G. reported the challenges of time self-teaching involved. There are many new skills associated with blog practice: make & post video / photo, archive and organize (e.g. tag) blog. Without official training, time is the biggest challenge for them to teach themselves required skills to make their blogs interesting and fun. They usually worked on their blogs after 8pm during their free family time.

4.3 Privacy Issue

Is your blog public or private? Do you have any policies on your classroom blog usage?

All three blog sites are public. Mr. B. did expressed concern about posting students pictures on his blog even when pictures were taken at school. He mainly posted student written work on his blog. In contract, Chinese teachers did not worry about the privacy issue much. When facing one class teacher to 60 families, their first priority in their communication with parents is the efficiency. For example, we observed that arguments or fights in the classroom were described on teacher's blog with student's names. Other students and parents would comment on those incidents and reflect on how to prevent them from happening in the future. Again this practice reflected the collectivism nature of Chinese culture. Sharing lessons within a group is more important than individual embarrassments.

5 Conclusion

The objective of this study is to find out how blogs, a lightweight web 2.0 technology, are used to support communication between parents and teachers in different national culture settings. The findings of this interview study suggested that cultural values, privacy policies, teacher background and technology knowledge have influenced the use of blog in parent-teacher communication.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. IIS-0803225.

References

1. Becker, H.J., Epstein, J.L.: Parent involvement: A survey of teacher practices. The Elementary School Journal 83(2), 85–102 (1982)
2. Center for Social Organization of Schools. Study of teacher practices of parent Involvement: Results from surveys of teachers and parents (Rep. No. 143). The Johns Hopkins University, Baltimore, Maryland (November 1983)
3. Epstein, J.L.: Parents reactions to teacher practices of parent involvement. The Elementary School Journal 83(3), 277–294 (1986)
4. Grolnick, W.S., Slowiaczek, M.L.: Parents involvement in childrens schooling: A multidimensional conceptualization and motivational model. Child Development 65(1), 237–252 (1994)
5. Kohl, G.O., Lengua, L.J., McMahon, R.J.: Parent involvement in school conceptualizing multiple dimensions and their relations with family demographic risk factors. Journal of School Psychology 38(6), 501–523 (2000)
6. Fan, X., Chen, M.: Parental involvement and students' academic achievement: A meta-analysis. Educational Psychology Review 13(1), 1–22 (2001)
7. Banach, W.J.: The abc's of teacher-parent communication. Rowman and Littlefield Education, Lanham, Maryland (2007)
8. Powell, D.: Correlates of parent-teacher communication frequency and diversity. Journal of Educational Research, 71(6), 333–341 (1978)
9. Halsey, P.: Parent involvement in junior high schools: A failure to communicate. American Secondary Education 34(1), 57–69 (2005)
10. Seitsinger, A.M., Felner, R.D., Brand, S., Burns, A.: A large-scale examination of the nature and efficacy of teachers' practices to engage parents: Assessment, parental contact, and student-level impact. Journal of School Psychology 46(4), 477–505 (2008)
11. Mitchell, S., Foulger, T.S., Wetzel, K.: Ten tips for involving families through internet-based communication. Young Children 64, 46–49 (2009)
12. Tobolka, D.: Connecting teachers and parents through the internet. Tech. Directions 66(5), 24–26 (2006)
13. Thompson, B.: Characteristics of parent-teacher e-mail communication. Communication Education 57(2), 201–223 (2008)
14. Kim, H.N.: The phenomenon of blogs and theoretical model of blog use in educational contexts. Computers Education 51(3), 1342–1352 (2008)
15. Richardson, W.: Blogs, wikis, podcasts, and other powerful web tools for classrooms. Corwin Press, Thousand Oaks (2006)

16. Zawilinski, L.: HOT Blogging: A framework for blogging to promote higher order thinking. The Reading Teacher 62(8), 650–661 (2009)
17. Lasky, S.: The cultural and emotional politics of teacher-parent interactions. Teaching and Teacher Education 16(8), 843–860 (2000)
18. Joshi, A., Eberly, J., Konzal, J.: Dialogue across cultures: teachers' perceptions about communication with diverse families. Multicultural Education 13(2), 11–15 (2005)
19. Huntsinger, C.S., Jose, P.E.: Parental involvement in children's schooling: Different meanings in different cultures. Early Childhood Research Quarterly (2009)
20. Snyder, T.D., Dillow, S.A.: Digest of Education Statistics (NCES 2010-2013). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC (2010)
21. Sparks, S.D.: Class-Size Limits Targeted for Cuts. Education Week 30(13), 1–16 (2010)

How Font Size and Tag Location Influence Chinese Perception of Tag Cloud?

Qiping Zhang[1], Weina Qu[2], and Li Wang[2]

[1] Palmer School of Library and Information Science, Long Island University
[2] State Key Laboratory of Brain and Cognitive Science, Institute of Psychology,
Chinese Academy of Sciences, Beijing 100101, China
qiping.zhang@liu.edu, {quwn,wangli}@psych.ac.cn

Abstract. Social tagging as a new approach for metadata creation has emerged to support browsing, searching, sharing on social network sites. Tag clouds are visual displays of social tags. In this paper we reported a user study on tag cloud perception. The goal of our evaluation is to investigate the effect of some of the different properties that can be utilized in presenting tags e.g. tag font size, tag location. Both behavior data and eye tracking data demonstrated a significant effect of font size, but effect of tag locations was mixed. Big tags were recalled better than medium and small font tags regardless of their locations in a tag cloud. Tags in the middle circle of a tag cloud received longer eye duration than outer circle, but were not recalled better.

Keywords: tag cloud, tagging, evaluation, visualization, user studies, Chinese.

1 Introduction

A tag cloud is a visual display of a set of words related to an information item such as a bookmark of a website, a blog entry, a photo. It usually has one purpose: to present a visual overview of the set of words. The size of a word is determined by the popularity to the tagged object. The larger the tag is, the more frequent the tag has appeared. In spite of their simple form, tagclouds have drawn a lot of attention from research communities [1,2, 3, 4, 5].

Several studies have done on the usage of tag clouds. Rivadeneira et al. [6] performed evaluation studies on searching and impression formation. As for goal-orientated tasks, simple alphabetical word lists are preferred over tagclouds [7]. Kaser et al. [8] proposed algorithms to create 2D tagclouds. In addition, studies concluded that the main value of tagclouds is as a signal or maker of individual or social interaction with information [5, 9].

Tags function as keywords, can be categorized with any word that defines a relationship between the online resource and a concept in the user's mind [10]. Tagging is implicitly also a social indexing process, since users share their tags and resources, constructing folksonomy, a social tag index, supporting visual information retrieval [11]. In regards to the issue of whether the tagcloud is actually useful as an aid to find information, Sinclair and Cardew-Hass's research found that where the information-seeking task was more general, participants preferred the tagcloud to

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 273–282, 2011.

search information [12]. Often, more frequently used tags are depicted in a larger font or otherwise emphasized. In essence, the tag cloud translates the emergent vocabulary of a folksonomy into a social navigation tool [13].

Considering the role of tag cloud as visual information retrieval, it is important to attract user's attention rapidly. However, perception of tag clouds is influenced by many attributes of tags such as font size, word orientation (horizontal vs. vertical) or color that are used to represent tag frequency or semantic relationship of tags. Given the conflicting results on variables like location, it remains unclear that how tag clouds are perceived visually and which search strategies users apply when looking for tags in a tag cloud. While most previous work were based on behavior data, we report a study on the perception of tag clouds using both behavior data and eye tracking data that allows answering these questions. The goal of the study is to investigate how font size and tag location influence Chinese perception of tag cloud.

1.1 Tag Cloud Design

To construct a tag cloud, we divide a square into 4 quadrants: upper-left, upper-right, lower-left, and lower-right. (Fig. 1)

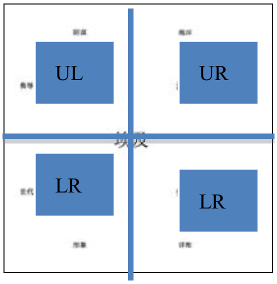


Fig. 1. Tag cloud Quadrants
(UL=Upper-Left; UR=Upper-Right; LL=Lower-Left; LR=Lower-Right)

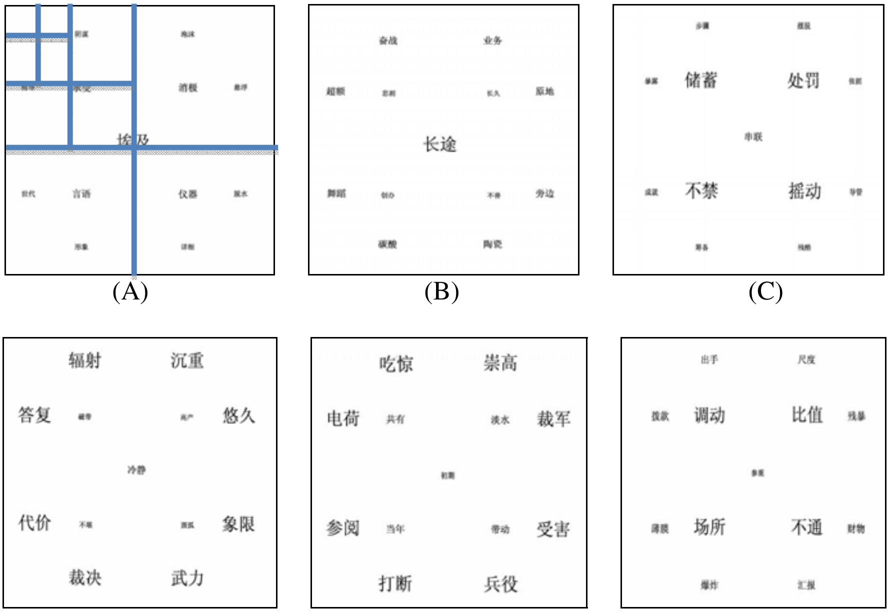
Then we constructed tag clouds by varying the tag font size and tag location based on the following rules:

a) Three font sizes were chosen based on previous study [6]: big-9.83mm, medium-5.42mm, small-1.38mm;
b) 3 tag locations in a tag cloud were defined by quadrantizing a square three times. Inner location is the center of the square. Only one tag is assigned in this location. Middle location is the center of each quadrant, thus 4 tags are assigned to this location. Within each quadrant, its outer sub-quadrant (e.g. upper-left sub-quadrant of upper-left quadrant, or lower-right sub-quadrant of lower-right quadrant etc.) was divided. Outer locations are the two intersection points of the sub-quadrants. Therefore a total of 8 tags are assigned to this location in a tag cloud. In total, there are 13 tags in each tag cloud (refers to A in Fig. 2).
c) Font sizes (big, medium, small) were counter balanced cross all 3 tag locations (inner, middle, outer).

In summary, there are 6 types of tag cloud combing tag location and tag font size (Table 1 & Fig. 2).

Table 1. Six types of tag clouds combining font size and location

	Tag Location		
Example	Inner (1)	Middle (4)	Outer (8)
A	Big	Medium	Small
B	Big	Small	Medium
C	Medium	Big	Small
D	Medium	Small	Big
E	Small	Medium	Big
F	Small	Big	Medium



(A) (B) (C)

(D) (E) (F)

Fig. 2. Examples of six types of tag cloud

2 Method

The experimental was a 3x3 repeated measures design, with font size (Big: 9.83mm, Medium: 5.42mm and Small: 3.94mm), Locations (inner, middle and outer) as independent variables. The behavior measure (recall) and eye tracking measure (eye fixation duration, and eye fixation times) were our dependent variables.

2.1 Participants

Twelve students (6 female, 6 male) whose ages ranged from 21 to 31, were paid 20yuan to participate. All subjects had normal or corrected-to-normal vision.

2.2 Material and Apparatus

First, a set of 403 2-character Chinese words was obtained from the Chinese Words Frequency Dictionary within the written frequency from 100 to 1000 (this frequency is the mode in this dictionary). This dictionary totally contained 31187 2-character words. Second, thirteen different words were randomly sampled from the set of 403 words, and displayed in predetermined locations in order to appear as a spatial tag cloud. The layout of the tag clouds is one word inner, four words middle and eight words outer. There are 6 types of tag cloud combing tag location and tag font size (Table 1 & Fig. 2). For each type, 5 tag clouds with another 13 different words were continually sampled from the set till a total of 30 tag clouds were formulated. The tag cloud was presented to the participants by eye movement equipment, Tobii 1750.

2.3 Procedure

Participants performed one practice trial and 30 experimental trials. Each trial began with a blank screen for 1s, followed by a tag cloud for 20s. And then, a distract task (participants had to count backwards in threes starting from a random number) followed for 30s in order to eliminate any recency effect. The trial ended with a 60-second free recall. The presentation orders of each type of tag cloud were counterbalanced among participants.

3 Results and Discussions

In the following, we reported our behavior data and eye tracking data.

3.1 Behavioral Data

A 3x3 repeated measure of ANOVA was conducted on the recall data. Not only the main effects of font size ($F(2, 22) = 46.57, p<.001$), and location ($F(2, 22) = 6.46, p < .05$) were significant, the interaction effect was significant ($F(4, 44) = 6.22, p < .001$) as well.

Table 2 and Fig. 3 showed the mean and standard errors of recall data.

Table 2. Recall of tags by font size and location (mean, standard error)

	Tag Location			
Font Size	Inner	Middle	Outer	Mean
Big	.66 (.05)	.39 (.03)	.35 (.03)	0.47
Medium	.29 (.06)	.22 (.03)	.27 (.03)	0.26
Small	.29 (.06)	.21 (.03)	.23(.03)	0.24
Mean	0.41	0.27	0.28	

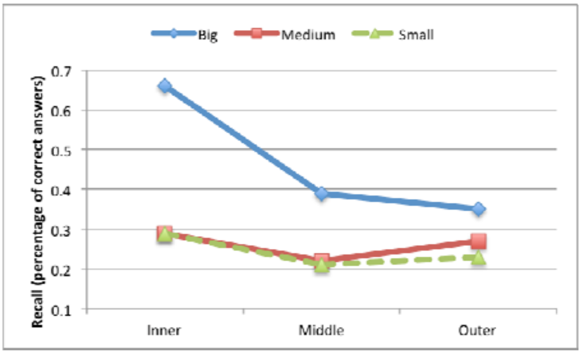


Fig. 3. Recall by font size and location

Recall for tags with a larger font size is significantly bigger than for tags with a smaller font size crossing all three locations. Pairwise comparisons revealed that when the location is inner or middle, the differences between big-medium, and big-small font size were significant, while the difference between medium and small font was not significant; when the location is outer, the differences among all pairs big-medium, big-small, medium-small were significant. This suggests that the advantage of medium font size vs. small font size on recall only occurs in the outer tag cloud. One possible explanation is that recall of tags in the middle tag cloud was interfered by tags from both inner and outer tag clouds, thus even out the font size advantage. A future study with different medium font size might further our understanding on this.

Recall for tags in different locations (inner, middle, outer) depends on the tag font size. When the font size is big, the effect of location is significant, F (2, 22) = 20.42, $p < .001$. Pairwise comparisons revealed that the differences between inner and middle (.66 vs. .39), and inner and outer (.66 vs. .35) were significant, while the difference between middle and outer was not significant. However, when the font size were medium and small, the effects of location were not significant. This result implies that font size seems a stronger factor than tag location in tag recall. This is an important finding, suggesting that designers should use the feature of tag font size rather than tag location to improve the recall and perception of tags in a tag cloud.

To further our understanding of effect of location on recall, one-way ANOVA of quadrant on recall was conducted. Main effect of quadrant was significant (F (3, 47) = 3.668, $p < .05$). As shown in Table 3 and Fig. 4, Recalls of tags in the upper quadrant (UL) were significantly better than the lower quadrants (LL, LR), but the difference between Upper-Left and Upper-Right (.34 vs. .33) was not significant. This confirmed our prediction that participants would pay more attention to upper quadrant than lower quadrant in a tag cloud given our reading habit is scanning from upper-down and left-right. In addition, our data suggested that better recalls of tags in a tag cloud only along the direction of upper-down, but not along left-right direction.

Table 3. Recall of tags by quadrant (mean, standard errors)

Quadrant		*Mean*	*SE*
Upper-Left	(UL)	.34	.042
Upper-Right	(UR)	.33	.042
Lower-Left	(LL)	.21	.021
Lower-Right	(LR)	.24	.015

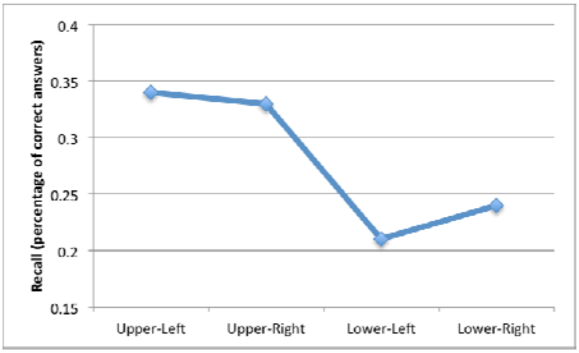


Fig. 4. Recall of tags by quadrant

3.2 Eye Movement Data

Similar to the behavior data, a 3x3 repeated measure of ANOVA with two within-subjects factors of three font sizes (big, medium, small) and three locations (inner, middle and outer) was conducted on the eye movement data.

First, the analysis of average fixation durations (seconds) revealed significant main effects of font size ($F(2, 18) = 14.92$, $p < .001$), and of location ($F(2, 18) = 3.71$, $p < .05$) as shown in Table 4 and Fig. 5.

Eye fixation duration for tags with a larger font size is significantly longer than for tags with a smaller font size as shown in the significant pairwise comparisons of all three pairs big-medium, big-small, medium-small ($p<.05$). Consistent with our behavior data, effect of font size is robust. This eye tracking data supported our behavior data that tags in bigger font size than in smaller font were scanned longer, processed deeper, and therefore recalled better.

In terms of effect of location on eye fixation duration, participants spent less time on tags in the outer tag cloud than in inner or middle tag cloud ($p<.05$), but the difference between inner and middle (353.79 ms vs. 346.76 ms) was not significant. This result supported our explanation to the behavior data that tags in the middle circle of a tag cloud received the interferences from tags in both inner circle and outer circles. Here participants spent similar amount of time on tags in the middle circle as those in the inner circle, implying that participants developed their own scanning strategy to the tags in the middle circle to reduce the interferences to them.

Table 4. Eye fixation duration (milliseconds) by font size and location (mean, standard error)

	Tag Location			
Font Size	Inner	Middle	Outer	Mean
Big	456.93 (70.75)	494.34 (67.47)	422.85 (63.66)	**458.04**
Medium	385.76 (81.54)	292.03 (49.31)	219.76 (46.72)	**299.18**
Small	218.68 (62.96)	253.90 (63.35)	178.39 (40.46)	**216.99**
Mean	**353.79**	**346.76**	**273.67**	

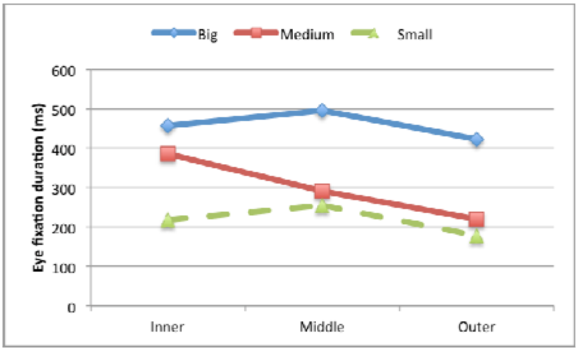


Fig. 5. Eye fixation duration (ms) by font size and location

Second, the analysis of average fixation times revealed a significant main effect of font size ($F(2, 18) = 29.43$, $p < .001$), a marginally significant main effect of location ($F(2, 18) = 3.54$, $p = .051$), and a marginally significant interaction effect ($F(4, 36) = 2.26$, $p = .081$) as shown in Table 5 and Fig. 6.

Similar to the result with eye fixation duration, eye fixation times for tags with a larger font size is significantly more frequent than for tags with a smaller font size as shown in the significant pairwise comparisons of all three pairs big-medium, big-small, medium-small (p<.05).

Different from eye fixation duration result, pairwise comparisons of locations revealed that only the difference between middle and outer was significant (1.29 vs. 1.00), while the differences between inner and middle, inner and outer were not significant. This further supported our explanation that participants developed corresponding visual scanning strategy on processing the tags in the middle circle of a tag cloud, that is scanning tags in the middle location longer and more frequently.

Table 5. Eye fixation times by font size and location (mean, standard error)

	Tag Location			
Font Size	Inner	Middle	Outer	Mean
Big	1.73 (.24)	1.97 (.23)	1.65 (.21)	**1.78**
Medium	1.22 (.20)	1.06 (.16)	.75 (.12)	**1.01**
Small	.59 (.13)	.86 (.16)	.59 (.09)	**.68**
Mean	**1.18**	**1.29**	**1.00**	

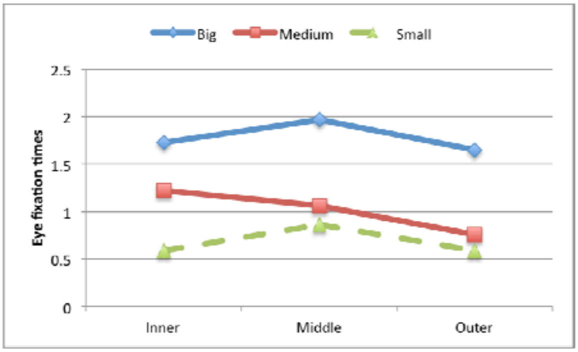


Fig. 6. Eye fixation times by font size and location

Finally, One-way ANOVA of effect of quadrant on eye movement data was conducted. It revealed a non-significant of quadrant on eye fixation duration (*F* (3, 39) = 1.817, *p* = .161), and a marginally significant effect on eye fixation times (*F* (3, 39) = 2.551, *p* = .071), as shown in Table 6 and Fig. 7.

Table 6. Eye fixation duration (seconds) and fixation times by quadrant

Quadrant		*Fixation duration*	*Fixation times*
Upper-Left	(UL)	94.45 (10.02)	365.80 (29.84)
Upper-Right	(UR)	87.70 (9.04)	335.60 (24.75)
Lower-Left	(LL)	72.17 (9.50)	288.60 (35.68)
Lower-Right	(LR)	66.92 (.49)	257.20 (29.73)

While participants did not spend longer time on upper quadrants than lower quadrants, they did focus their eyes more frequently on upper quadrants than lower quadrants. Post Hoc Tests revealed significant difference of eye fixation times between UL and LR (365.8 vs. 257.2) was significant, and UL and LL (365.8 vs.288.6), UR and LR (335.6 vs. 257.2) were marginally significant. Again, this results suggested participants focused their eyes more frequently in the upper than lower quadrants, and more in the left than in the right quadrants.

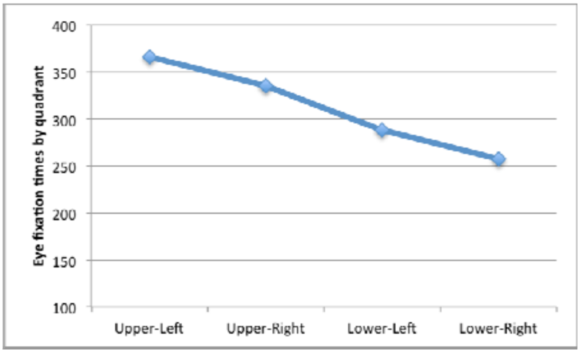


Fig. 7. Eye fixation times by quadrant

4 Conclusion and Implications

Overall, both our behavior data and eye tracking data demonstrated a robust effect of font size. Participants scanned and recalled more tags with larger fonts. They recalled more big tags than medium and small tags, while no more medium tags than small tags, except when the location is outer. However, we found a significant difference between medium tags and small tags in eye movement. Participants spent longer time and switched their eyes more frequently to medium tags than small tags, but they didn't encode or retrieve more medium tags. Therefore, eye movement is a more sensitive index to indicate what kind of tags can attract people's attention. People browse and search the tag clouds, click the tags that they are interested in, and needn't to memorize the tags.

The effect of location is not as robust as that of font size, and the behavioral data and the eye movement data are not so consistent. Only when the tags are big, participants recalled more tags inner than middle and outer. Interestingly, participants seemed developed their own visual scanning strategy to the tags in middle of a tag cloud. They spent similar amount of time, on tags in the middle as in the inner location, but significantly more time than in the outer location. The eye fixation frequency showed similar findings. These findings suggested while participants tried to move their eyes to tags in the middle location to reduce the interferences from tags from outer and inner locations, the recall of tags in the middle location was not compensated.

The effect of quadrant showed that participants recalled more tags in upper-left and upper-right than lower-left and lower-right and scanned more in upper-left than lower-right. Designers may consider upper quadrant especially upper-left as a focal point within a tag cloud. This area can either locate smaller font tags to compensate for font size, while locating bigger font tags to other quadrants, or locate tags that need to be emphasized.

In the future, other features of tag clouds such as tag colors, tag orientation, tag cloud size/density, number of characters/tag, semantic relationship of tags, need to be investigated so that we will have a better understanding of both cognitive and social process of tags in a tag cloud in order to design a better retrieval system to support tag creation, presentation and sharing.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. IIS-0803225. We thank Dr. Kan Zhang and Dr. Xianghong Sun for their assistance in this project.

References

1. Eda, T., Uchiyama, T., Uchiyama, T., Yoshikawa, M.: Signaling emotion in tagclouds. In: Proceedings of the 18th international conference on World wide web, Madrid, Spain, April 20–24 (2009)
2. Schrammel, J., Deutsch, S., Tscheligi, M.: Visual Search Strategies of Tag Clouds - Results from an Eyetracking Study. In: Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part II, Uppsala, Sweden, August 24–28 (2009)

3. Chen, Y.-X., Santamaría, R., Butz, A., Therón, R.: TagClusters: Semantic Aggregation of Collaborative Tags beyond Tag clouds. In: Proceedings of the 10th International Symposium on Smart Graphics, Salamanca, Spain, May 28–30 (2009)
4. Cheng, J., Cosley, D.: kultagg: ludic design for tagging interfaces. In: Proceedings of the 16th ACM international conference on Supporting group work, Sanibel Island, 07-10, Florida, USA, November 7–10 (2010)
5. Hearst, M.A., Rosner, D.: Tag clouds: Data analysis tool or social signaller? In: Proceedings of HICSS (2008)
6. Rivadeneira, A.W., Gruen, M.D., Muller, J.M., Millen, R.D.: Getting Our Head in the Clouds: Toward Evaluation Studies of Tagclouds. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 995–998 (2007)
7. Halvey, M.J., Keane, M.T.: An assessment of tag presentation techniques. In: Martin, J., Halvey, M.T. (eds.) Proceedings of the 16th international conference on World Wide Web, May 08-12, Banff, Alberta, Canada (2007)
8. Kaser, O., Lemire, D.: Tag-cloud drawing: Algorithms for cloud visualization. In: Proceedings of WWW (2007)
9. Viégas, F.B., Wattenberg, M.: Tag clouds and the case for vernacular visualization. Interactions 15(4), 49–52 (2008)
10. Guy, M., Tonkin, E.: Folksonomies: tidying up tags? D-Lib Magazine, December 1 (2006)
11. Hassan-Montero, Y., Herrero-Solana, V.: Improving tag-clouds as visual information retrieval interfaces. In: Proc. InfoSciT (2006)
12. Sinclair, J., Cardew-Hall, M.: The folksonomy tag cloud: when is it useful? Journal of Information Science. Journal of Information Science 34(1), 15–29 (2008)
13. Dieberger, A., Dourish, P., Höök, K., Resnick, P., Wexelblat, A.: Social navigation: techniques for building more usable systems. Interactions 7(6), 36–45 (2000)

Part IV
Cognition and Automation

Balance between Abstract Principles and Concrete Instances in Knowledge Communication

Toshiya Akasaka and Yusaku Okada

Keio University, Faculty of Science and Engineering, 3-14-1 Hiyoshi, Kohoku-ku,
Yokohama, Japan
{to48_a,okada}@ae.keio.ac.jp

Abstract. In tasks requiring dealing with variable situations, workers are expected to do more than following prescribed instructions. In this paper, we presented our view and framework for creating instructions with a good balance between the flexibility of abstract principles and the preciseness of concrete instances, which aims at helping instruction receivers become capable of dealing with variable situations where no concrete instructions are available. Our approach represents knowledge using an abstraction hierarchy. It is situated in our grand model which deals with the whole picture of knowledge communication. A case study is also presented, which suggests that seen in our view existing manuals can be improved by providing them with principles combined with instance-dependent variables.

Keywords: Hierarchic Representation of Knowledge, Modeling Language, Knowledge Management, Knowledge Visualization.

1 Introduction

Teaching how to perform a task is always difficult, especially when the task process is complicated. Usually, instructions are needed to communicate required knowledge. Sometimes, instructions take the form of written manuals and are provided for such tasks as operation of man-made systems like plant operation, operational business activities like call center, management systems implementation like project management, and so forth. Manuals of these tasks are used not only to help new comers know how to perform the task, but also to share common knowledge among people involved in the task.

Manuals are written with the aim of giving explicit instructions free of ambiguity. For example, some manuals about plant operation describe the sequences of required operations in a well-structured form, defining contexts in a concrete manner and specifying exact buttons and levers to operate for each context as well as the order of those operations. Aiming to avoid ambiguity, such manuals sometimes end up being the compilation of case-by-case operation sequences, still falling short of covering all possibilities. In other words, manuals fail to convey principles and full range of possibilities as a result of its aim to take a rigorous, ambiguity-free form.

Receivers of instructions taking such a rigorous form often have difficulty with having well-organized understandings of the task; they simply follow prescribed instructions and do not think of any principle from which such concrete instructions

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 285–293, 2011.

are derived. This means that they cannot take appropriate actions without instructions. Unfortunately, there are many cases for which sufficient instructions are not available, such as troubleshooting in software applications, incident handling in plant operation, demanding-customer handling in call center, and so forth. Facing novel, unfamiliar situations, they are at a loss what to do, even if the situations can be resolved by making use of the actions they are used to. They know fragmentary pieces of knowledge, but do not know what of them can be applied to the actual situation in what way.

In work settings, it is important for workers to be capable of dealing with such unfamiliar situations. If they otherwise could not deal with situations, appropriate performance would be hard to guarantee. Also, if they always consult well-experienced people every time they fail to find explicit instructions, such experienced workers may find it difficult to concentrate on their own work. This is all the more problematic if veteran workers are decreasing in number. Of course, it is almost impossible to make novices be experts simply by providing instructions, however good they may be. Still, giving them a bid of flexibility makes a big difference if readers otherwise could do no more than following explicit instructions. To succeed in this, there needs to be a good medium of knowledge, giving readers the flexibility to take appropriate actions for variable situations.

However, giving such flexible instructions is not an easy task. The challenge here is how to balance the flexibility of abstract principles and the preciseness of concrete instances. Obviously, describing all alternatives in minute detail is impossible, since such detail instructions are infinite in number as is the case with troubleshooting of software errors. On the other hand, if one tries to describe instructions in a larger granularity, the number of alternatives may become of a manageable magnitude, but there is a risk of causing instruction receivers to be at a loss what to do due to the loss of concrete instructions.

The goal of our research is to investigate a systematic way to provide instructions in the way that gives principled explanations in an abstract description while covering an appropriate variety in a concrete description. In this paper, we present our research approach toward tackling this challege as well as the current state of progress in discussion and investigation.

2 Research Approach

2.1 Abstraction Hierarchy

Our approach involves representing knowledge using an abstraction hierarchy. If some instances are derived from a certain model, rule, law, or any other form of general principle, that principle is said to be abstract as compared to those instances, while the instances are said to be concrete as compared to the principle. Likewise, if some means are to achieve a certain goal, the goal is abstract as compared to those means, and the means are concrete as compared to the goal. In short, something abstract is the reason why something concrete appears or is used. Consequently, we can find more concrete instances than abstract ones, for from one abstract principle or end come several (sometimes infinite) instances or alternative means. The abstraction hierarchy considered here has several layers each representing a certain degree of abstraction, and elements that are considered to be abstract to the same degree are situated on the same layer. Each layer is abstract to the lower one and concrete to the upper one. Thus, we can find more instances as we go down along the hierarchy. The illustration of our abstraction figure is shown in Figure 1. Note that this is just a rough sketch of our idea, just intended to help readers better understand it. We greatly owe this abstraction hierarchy to Rasmussen [1]. Rasmussen, who had long studied human-machine interface in major physical systems such as plant operation systems, discovered that operators of such systems seemed to have several information processing modes; they first processed a set of numerical data on the information display, and the next moment they thought of the system structure including mass flow model, a totally different type of thinking from recognizing numerical data. He then went on to propose a structured representation of a physical system as it appeared to operators, with the aim of developing a framework to describe human-machine interaction. Thus, he presented an abstraction hierarchy with each layer representing the structure of a physical system from the viewpoint of a certain abstraction level. Although limited to physical systems structure, Rasmussen's idea of representing human's information processing as a travel along the abstraction hierarchy is a generally applicable framework for studying knowledge and its explanations.

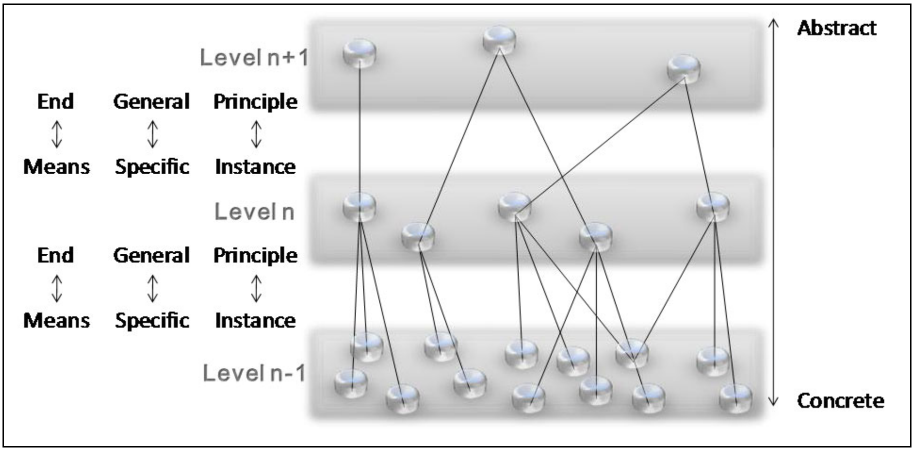


Fig. 1. Illustrative sketch of abstraction hierarchy. As going up along the hierarchy, one needs fewer descriptions to explain the same range of possibilities (illustrated by the number of circles), while description themselves become less self-explanatory, demanding supplementary explanations.

Using the abstraction hierarchy, we can represent our problem differently. First, we can say that if somehow organized, instructions are consistent in terms of abstraction level. Most instructions are given in a specific abstraction level. Take manuals of software applications for example. Some manuals give instructions by sequences of GUI (Graphical User Interface) level operations (e.g. "Select the file item on the menu bar"), while others, perhaps those for professional software engineers, may use sequences of more granular operations like "Save the document in .txt format", "Install the driver software with the bundled CD-ROM." It is fair to say that the instructions of the second type are unlikely to coexist with those of the first type and vice versa, probably due to the difference in intended readers. The abstraction level differs across manuals, but mostly remains consistent within one.

Given that a set of instructions are provided in a consistent level of abstraction that is low enough for intended readers to comprehend the instructions (let us call this level the *ground level*), then the next step is how to help readers become capable of dealing with unexpected situations. If one tries to explain all possible solutions to every possible situation in the ground level, it may be infeasible due to an unacceptably large number of possibilities; the writer may not be able to recognize all possibilities and find sufficient pages to write them down on. A possible solution here is to raise the level of abstraction. Remember that from one abstract principle or end come several (sometimes infinite) instances or alternative means. Translating this, we can have that with a higher level of abstraction you can explain a broad range of possibilities with fewer descriptions. The problem here is how to balance between the flexibility of abstract principles and the preciseness of concrete instances. Writers are now demanded to explain a principle while showing a variety of possibilities derived from that principle, both in a comprehensible manner. On top of that, the description has to be smarter than simple enumeration of all possibilities at the ground level of abstraction; it is meaningless to make use of higher-level principles unless the resulting description becomes easier to comprehend than the original one. Thus, our core problem can be formulated as follows; to investigate how to balance between the flexibility of abstract principles and the preciseness of concrete instances in a comprehensible manner, with the aim of helping instruction receivers become more capable of handling situations with no explicit instructions prescribed.

In our view, many of existing written manuals tend to describe only one or a few typical instances at the ground level of abstraction, and that tendency is even clearer for the manuals of large systems, like plant operation systems. Academically- and industrially-proposed modeling languages have also failed to show a promise. For example, there are many modeling languages intended for software systems development, from traditional ones like flowchart and dataflow diagram to such modern techniques like UML (Unified Modeling Language). They are fairly good at describing well-defined structures, but lack the flexibility to describe the variety seen in real-world tasks, demanding the modelers to make every instruction explicit. In fact, some professional developers already gave up using a modeling method as means of eliciting requirements, and instead have tried to ensure effective communication by rapid trial-and-error; they repeat a small cycle of producing a trial product and demonstrating it to the clients. After all, the modeling languages for software development need to be compatible with computer-oriented digital processes, and it is only natural that they cannot tolerate ill-defined possibilities in the real world. On the other hand, there is a modeling language which intends to describe human-involved processes. The functional modeling language IDEF0 [2], which was based on the SA (Structured Analysis) language [3], a graphical modeling language, has as its premise the belief that describing and managing human activities is important in order to make good use of information technology systems. Like our concept, the SA language features the hierarchic structure of a subject. Although its rigorous grammar does not immediately lead to flexible expressions, the SA language provides users with unique notations which, together with the hierarchic structure, successfully represent reasons and mechanisms. Careful study and some tailor of the SA language may bring us a promising method to express both abstract principles and concrete instances.

2.2 Grand Model

Although our current problem has been already delineated, we now briefly sketch our grand model to situate our current investigation in the whole picture. We have been discussing how an instruction should be given to its receiver. Put in a broader perspective, those who give instructions also used to be novices in the past. They learned a lot by first-hand experience and instructions from yet other people. That is why they are now capable of handling variable situations; in their mind are lots of instances and abstract principles. To give instructions is to decode such abstract principles and transform them into visible description with the help of memories regarding past instances. Then, the resulting instructions convey knowledge to instruction receivers. However, such an easy story is hardly the case. Unfortunately, the fact is that experience is the hardest thing to communicate and the gap between well- and poor-experienced people has been never seen bridged.

One possible reason is that some parts of experience are filtered out, not taken into account when abstraction occurs. Having experienced many instances and learning lots of pieces of knowledge, people for some reason try to generalize them and find some abstract principles. Although it is impossible to describe the process of abstraction, many would agree that the process is not as simple as a strict one-way sequence. Instead, it should be a kind of parallel process; from a set of instances comes an incomplete principle which has not yet taken a complete form, while the instances are sought that fit the fetus principle well. This means that there can be some selections of instances where those which do not fit the incomplete principle are neglected. Sometimes, it is useful to allow different principles to coexist so as to absorb as many instances as possible, but that is not so easy. When instances and potential unknown principles are neglected in order to allow only one principle, the variety of concrete instances are likely to be neglected when instructions are given.

Another possible reason is that the expressive power of a language prevents principles and instances from being described properly. The type of language affects people's thinking. When writing computer program code, no one would think how to write an analogy of a certain algorithm, because there is no way to describe an analogy using programming languages. On the other hand, usual people do not think of class or instance when writing a natural language. This is also because natural languages are not good at describing such concepts. In other words, devising a good language can bring us a solution to communication problems. That is why many languages have been studied and some of them have been used widely.

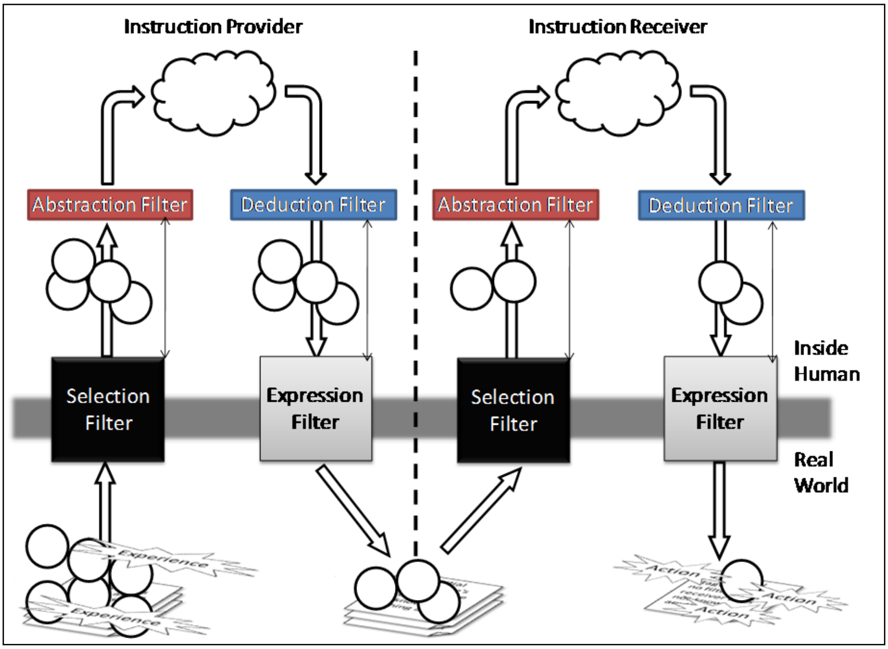


Fig. 2. Illustrative sketch of grand model

Yet another factor impeding communication is the gap of knowledge between provider and receiver of instructions. As we said earlier, decent instructions are given in a consistent level of abstraction and that level is decided so that the intended receivers of the instructions can easily comprehend what the instructions tell them. It is obvious that a problem emerges when that is not the case. Consequently, communication becomes more difficult when the gap is larger between the level that intended receivers can understand and the one that the target principle belongs to. Also, it follows that the higher level of abstraction the intended receiver can understand, the simpler instructions are.

Having discussed reasons for the difficulty in knowledge communication, we are now in the position to present the sketch of our grand model as shown in Figure 2. The instruction provider obtains abstract principles based on filtered experience and knowledge. Abstraction and selection are interrelated to each other. When trying to describe instructions, the provider has to deduce possible instances from principles and describe them along with the principles. If s/he fails to recognize important instances, the resulting instructions will have a potential failure however expressive the description language is. The receiver of instruction receives the instructions and follows the same route as the provider, although there is no guarantee that the receiver understands the instructions as the provider intends. If the instruction receiver in the figure is to give instructions to yet another person, s/he is now a provider of instructions and has to give instructions instead of taking actions, continuing information relay. This is how an organization is managed and its strategy instructed by top managers is disseminated until finally implemented. The figure shows four types of filters, namely, selection, abstraction, deduction, and expression filters, each of which can be the cause of a communication failure. Some of the failures have been already mentioned, but the entire list is presented here;

Neglect of Instances; receiver who has an incomplete principle still under construction neglects the instances that do not fit the principle well (interrelated to *neglect of principles*).

Neglect of Principles; receiver who already had one principle, complete or not, do not create another one even though there are instances that cannot be explained fully by the existing principle (interrelated to *neglect of instances*).

Failure of Abstraction; receiver recognizes instances but cannot find any principle behind them.

Failure of Deduction; provider who has a relevant principle fails to provide some important instances derived from that principle.

Failure of Expression; provider who has a set of relevant instances in mind fails to express all of them.

3 Case Study

In this section, we investigate a practical manual and see how it can be improved according to our view. The manual used here is the one for a power plant operation. It consists of operation manuals for start-up, shutdown, and emergency situations. The emergency situations part consists of case-by-case references each indicating a sequence of operation. The start-up and shutdown situations parts are further divided into about 40 steps; still each step includes tens of events and operations. Each operation indicates the exact button to operate or measure to observe. Although each step has the label indicating what the step is like in terms of function, within each step is just a sequence of operations, describing neither principles nor alternatives. When studying in detail, however, one can find that each step consists of modular-like patterns of operations, such as pump starting-up pattern, valve closing pattern, pressure raising pattern, and so on. Sets of operations that seem to be an instance of a certain pattern often appear in several steps, but they are different in minute detail; a certain operation seen in one instance is absent from another instance; the order of operations is often slightly different from instance to instance. Of course, in the manual there is no description about to what extent these instances can be explained by a single principle. If there is a principle that can explain most of the instances, we can give principled explanations by combining the principle and instance-dependent variables. That will eventually help readers become more capable of dealing with variable situations since now they can learn the visualized principle. Thus, we performed a trial analysis focusing on the pump start-up pattern.

We investigated seven instances of the pump start-up pattern. Each instance is a sequence of operation on a different pump. The seven pumps are from the four categories which are summarized in Table 1, and the sequences of operations for all the seven pumps are shown in Table 2.

Table 1. Seven pumps categorized by the combination of pressure and type

	Type X	Type Y

High Pressure	HPX(A)
	HPX(B)
		HPY(A)
		HPY(B)

Low Pressure	LPX(A)	LPY(A)
		LPY(B)

Table 2. Sequences of operations for seven instances. A check mark indicates that the operation is required for that instance, while a dash mark indicates that the operation is missing from that instance. The sign *confirm* means that it is required to confirm that the target state of that operation is achieved (no action is required). Finally, the numbers indicate that the operation there requires to confirm that a certain measure equals the number.

	LPX(A)	HPX(A)	HPX(B)	LPY(A)	LPY(B)	HPY(A)	HPY(B)
Close Valve A	✓	*Confirm*	✓	*confirm*	*confirm*	*confirm*	*confirm*
Open Valve B	–	✓	✓	✓	✓	✓	✓
Start-up Pump	✓	✓	✓	✓	✓	✓	✓
Pressure (kg/cm^2)	18	50	47	26	26	59	59
Flow (m^3/h)	–	–	–	85	85	250	250
Check Water Level	✓	–	–	–	–	–	–
Onsite Inspection	✓	✓	✓	✓	✓	✓	✓
Valve B→Auto	–	✓	✓	✓	✓	✓	✓
Close Valve B	–	*Confirm*	*confirm*	*confirm*	*confirm*	*confirm*	*confirm*
Check Sample Water	–	–	–	✓	✓	✓	✓
Open Valve A	✓	✓	✓	✓	✓	✓	✓
Check Tank Level	–	–	–	✓	✓	✓	✓
Stop Sub-pump	–	✓	✓	–	–	✓	✓
Next Pump→Auto	✓	✓	✓	✓		✓	
Open standby valve	✓	–	✓	✓		✓	

As shown in Table 2, some operations are seen common in all instances, showing that there are invariant structures. We can also find high-pressure-specific and type-Y-specific invariant structures (e.g., "Stop Sub-pump" and "Flow"). Actually, variables that are thought to belong to no invariant structure do exist but only to a small extent. This suggests that a principled explanation is a feasible option. Probably, a possible principle in this case takes the form of hierarchy in which there are general pump category at the top layer, the high-pressure and low-pressure pump categories at the middle layer, and the four categories of the matrix in Table 1 at the bottom layer. With this principle, we would have three types of general rules each corresponding to one of the three layers. The top layer rule is a generalized version of the middle layer rule, which in turn results from generalizing the bottom layer rule. Then, the description of pump operations would be of the form "a general rule plus instance-dependent variables."

4 Conclusion and Future Work

In this paper, we presented our view and framework for balancing between the flexibility of abstract principles and the preciseness of concrete instances, which aims at helping instruction receivers become capable of dealing with variable situations where no explicit instructions are available. Our investigation is still in the process of trial analysis where we analyze various instructions in a case-by-case manner. By conducting more case studies, we hope that our model will become well-formulated and will be capable of guiding instruction design systematically.

References

1. Rasmussen, J.: Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering. Elsevier Science Publishing Co., Inc., New York (1986)
2. Knowledge Based Systems, Inc., `http://www.idef.com`
3. Ross, D.T.: Structured Analysis (SA): A Language for Communicating Ideas. IEEE Software Transaction Engineering SE3((1), 16–34 (1977)

Using Uncertainty to Inform Information Sufficiency in Decision Making

Xiao Dong[1] and Caroline C. Hayes[2]

[1] Industrial & System Engineering Department, University of Minnesota
[2] Mechanical Engineering Department, University of Minnesota
Minneapolis, MN, USA
dongx080@umn.edu, hayes037@umn.com

Abstract. Decision making is a critical part of design. Designers must constantly compare, weigh and select design options throughout the design process. The effectiveness of those decisions impacts the effectiveness of the final design. In this paper, we compare two decision support systems, one that allows designers to enter and visualize the uncertainty in each alternative, and one that does not. We compared differences in the designers' perceptions of whether they had sufficient information to make a choice, and their confidence in their choice. The goal is *not* to make designers more confident of their decisions, but rather to help them evaluate realistically whether they have sufficient information to make a clear choice.

Keywords: Decision support system, decision making under uncertainty.

1 Introduction

The goals of this work are to develop and evaluate a decision support system (DSS) which helps decision makers to more accurately identify situations in which they need to gather more information before they can make a choice between several alternatives with confidence. Many decision methods, and computer tools that implement those methods, focus on helping designers to make the best decisions possible. However, in any complex decision with important consequences, many important facets of the decision situation are either unknown or uncertain. In many cases, the information may not yet be known that will allow the decision maker to know which is the best choice. Thus in this work, we have focused on helping decision makers to better assess whether or not they have enough information to make a clear choice though a simple visualization inspired by sensitivity analysis. We hope that by helping decision makers to recognize when they lack information, it will encourage them to seek clarifying information. We have focused our evaluation on the domain of product design, and have used designers to assess the tool. However, we believe the results to be applicable to almost any type of decision making.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 294–302, 2011.

2 Literature Review

Engineering design is a complex and ill-defined task in which designers must make decisions even when critical information is unknown or uncertain [1, 2, 3, 4, 5]. Thousands of decisions must be made in the course of a design project, and the financial and other consequences of bad decisions can be catastrophic. Rational decision methods of various sorts [6] are often used by designers to improve decision making, especially for major decisions. For example, designers often create decision matrices on paper, in spreadsheets, or with the assistance of a DSS.

However, decision makers often find such methods problematic in several ways. It is inconvenient to enter all the information required for such a method [7]. Erev and Bornstein [8] found that simply allowing designers use DSS does not necessary to increase the quality of the design. Other studies [9, 10] investigated whether allowing designers to express uncertainty in design values would improve the quality of design choices. Not only did this study find that it did not have much impact on the quality of the final decisions, expressing uncertainty information also required significant additional time. Furthermore, Hayes and Akhavi [9] observed that entry-level designers did not always know when to seek more information or what information would be most valuable in clarifying a choice. Based on these results and observations, we came to the conclusion that helping decision makers to choose the "best" option may not necessarily make sense; given the lack of knowledge about the options on the table, in many situations it may be impossible to know which choice is the best, even with the help of the best tool. Thus, in the work reported in this paper, we have changed our focus from directly helping the decision maker to identify the "best" option, to helping decision makers identify whether they have enough information to make a choice. We hope that by focusing on information sufficiency, it will encourage decision makers to work more on gathering appropriate information, which will likely lead to better decisions in the end.

3 Methodology

3.1 Two Decision Support Systems

We developed a DSS tool that allows decision makers to specify their estimates of the uncertainty in each of design parameter, and to visualize the uncertainty in the overall suitability of each design. The top-level interface is shown in Figure 1, on the left hand side. In order to allow us assess the impact of the DSS and its visualization, we also created a second DSS which does not allow users to express or visualize uncertainty in the design parameters. This interface is shown on the right hand side of Figure 1. To distinguish the two interfaces, we will call the first one the "uncertainty" DSS, and the other the "certainty" DSS.

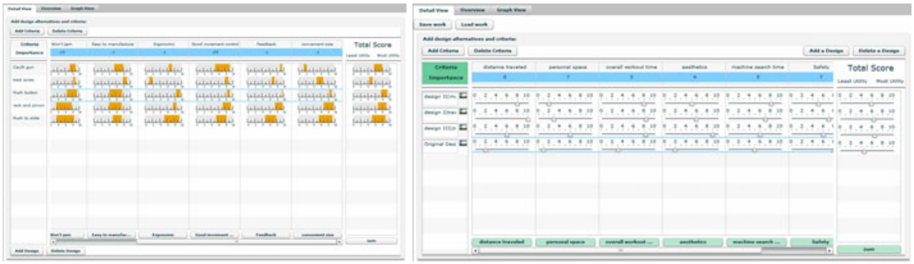


Fig. 1. Interfaces for two DSSs. On the left is the interface for the "uncertainty" DSS, on the right is the interface for the "certainty" DSS.

Both interfaces are structured like traditional weighted decision matrices, which are already familiar to many designers. Designers can enter the name of a design alternative at the start of each row. They can enter the names of their criteria at the top of each column, along with a number representing the weight or importance of that criterion. Each cell in the matrix created by these rows and columns represents a value indicating the degree to which that design alternative fulfills that criterion.

The "certainty" DSS allows users to enter only one value for each design parameter, and the overall scores are also displayed as single "point" values, as shown in the interface on the right-hand side of Figure 1. However, the "uncertainty" DSS allows users to enter values as ranges, using a pair of sliders. Thus, if the designer is uncertain about what exact value to enter, for example for the manufacturing cost, he or she can enter a range of values. The distance between the sliders (e.g. the width of the bar) indicates how much uncertainty is associated with each value.

Similarly, the "uncertainty" DSS computes an overall value score for each alternative (a weighted sum of minimum and maximum values), and is displayed in the right-hand column as a bar representing a range of values. Sometimes, one alternative stands far above the others. But more often, several "best" alternative may "overlap" in their overall value, as shown in Figure 1; the overall value scores of the third, fourth and fifth design alternatives are "better" than the other alternatives, but since their ranges overlap, it is not possible to tell which is really the best alternative. To clarify which is really the best, it is necessary to gather more information about the criteria that most contribute to the uncertainty in those alternatives.

By providing this simple set of bars to visualize the uncertainty in each alternative, designers can tell at a glance if there is one clear "winner" – e.g. a bar with the highest value and no overlap with the others, or whether there are several contenders -- indicating by overlapping bars. It is designed to be both simple and familiar so that users can learn to use it with relatively little training.

3.2 Research Questions

In this research, we wanted to investigate the following questions. 1) Can uncertainty visualizations be used to give designers a more accurate awareness of information sufficiency (e.g. whether or not they have sufficient information to evaluate which choice is best)? 2) Can uncertainty visualizations help designers to have more realistic confidence in their decisions (i.e. to be less confident when there is not enough information to make a clear choice)? 3) Can uncertainty information be used to encourage designers to seek clarifying information when appropriate? Finally, 4) Does domain experience change the benefits that decision makers derive from the DSS?

3.3 Experimental Design

We used a 2 x 3 within subjects design. The independent variables were *expertise level* (entry-level or intermediate-level designer), and *Design Comparison Method* (control system, "certainty" DSS, or "uncertainty" DSS). Dependant variables were perceived information sufficiency, effort to reach a decision, decision confidence, plans to seek additional information, and preference between the methods. Our hypothesis was that the "uncertainty" DSS would appropriately reduce designers' perceived information sufficiency and decision confidence when there were "overlaps" between the top alternatives, and that the others would not. In other words, it would help designers to notice when more information was needed before they could make a choice between alternatives with any confidence.

Subjects. We recruited 22 designers from mechanical engineering and medical device design backgrounds. Of these, 12 were entry-level designers (senior undergrads and one junior) and 10 were intermediate-level designers (graduate students from mechanical engineering department and center for medical devices). Entry-level designers had on average 1.5 years (SD = 0.84) of design experience, while intermediate-level designers had 3.1 years (SD = 1.22). The entry-level designers were on average 25.83 years old (SD=6.10), while intermediate-level designers were on average 31.80 years old (SD=4.87). All participants were currently working on design projects which had been underway for at least 4 weeks.

Design Tasks. Instead of giving each subject the same set of design tasks, we asked each subject to compare several options currently under consideration in their own on-going design projects. The reason was that we wanted to observe the impact in tasks that were both real and complex. It was also important that the subjects had had time to become reasonably knowledgeable about the specific design projects, and were highly vested in the decision outcome. None of that would be possible to recreate using and artificial "laboratory" design task.

Design Comparison Methods. Each subject was asked to compared different sets of design alternatives of their own choosing, using 3 different methods: using their normal practices (control condition), the "certainty" DSS, and using the "uncertainty" DSS. In the control condition subjects were allowed to use whatever methods or tools they would normally use to compare design alternatives. In some cases this was pencil and paper, and in others, a spreadsheet. The order in which they used the DSSs was systematically varied, so as to counter-balance learning effects.

Procedure. Each participant completed all tasks in the experiment individually. Participants were randomly assigned to one of the two experiment groups. The difference between groups was the order in which they used the "certainty" and "uncertainty" DSSs. Both groups used their normal method (control condition) first, then one group used the "certainty" DSS followed but the "uncertainty" DSS, while the other group did the reverse.

1. At the beginning of each experiment, participants were given a brief introduction to the study. Participants signed a consent form which included agreement to audio recording, and answered questions on their demographics and design experience.
2. Subjects were given a brief training session on both DSSs. The training sessions took roughly 10 minutes on each system (20 minutes total).
3. Participants were to use the three different methods in the order specified by the experimenter for their experimental group. They were asked to use the methods to identify one design from a set of alternatives which they would choose further development. Participants were given no time limit for completing design tasks.
4. Immediately after using each system, participants were asked to complete a questionnaire measuring all dependent variables, except "system preference".
5. The "system preference" question was answered after completion of tasks on all decision support systems.
6. At the end of the session, after using the three methods, participants were asked to reproduce the tasks they used in the "certainty" and control condition on "uncertainty" system. The purpose was to find out the degree of uncertainty (e.g. overlap) that existed between those alternatives.
7. A week later, the participants were asked to report what information they had gathered for their design projects.

3.4 Data Collected

Preference between methods was measured by asking subjects to "Please split 100 points among the three decision methods you have just used. The most preferred system should get the most points." *Effort to use decision support system*, *decision confidence*, and *perceived information sufficiency* were measured based on subjects answers to the statements: "I feel this method required more effort than should be necessary." "I am not very confident that the design(s) I chose were the best ones." "I feel I had sufficient information to make an informed decision." The answers to these questions were marked on 7-point Likert scales from "*Strongly Disagree*" (scored as 1) to "*Strongly Agree*" (scored as 7). *Plans to seek additional information* were assessed by first asking, "Is there a particular aspect of the design(s) on which you would like to know more in order to be confident of your decision?" and "If so, please elaborate below and indicate how strongly you want to know about it." The degree of desire was measured by a 7-point Likert scales ranging from "*Strongly Undesired*" (scored as 1) to "*Strongly Desired*" (scored as 7).

4 Results and Discussions

Method Preference. A repeated measure ANOVA, with expertise-level as between subject variable, was used to analyze the preference of the systems and effort to use the systems. We found a significant main effect on design comparison method

($F_{(2,40)}$ = 309.430, p < 0.001). As shown in Figure 2. There was also a significant interaction effect between design experience and decision support system ($F_{(2,40)}$ = 4.726, p = 0.014). These results indicate that both entry-level and intermediate-level designers preferred the "uncertainty" DSS much more than certainty DSS and control. This preference was stronger for the designers more experience, the intermediate-level designers.

Effort Required. Figure 2 shows that Entry-level designers found that it required significantly less effort to use the "certainty" DSS than their own method (i.e. no DSS) ($F_{(2, 22)}$ = 21.020, p < 0.001), but no significant difference between the two DSSs. For the intermediate-level designers, there was no significantly difference in the effort required to use the three different design comparison methods (control, and two DSSs).

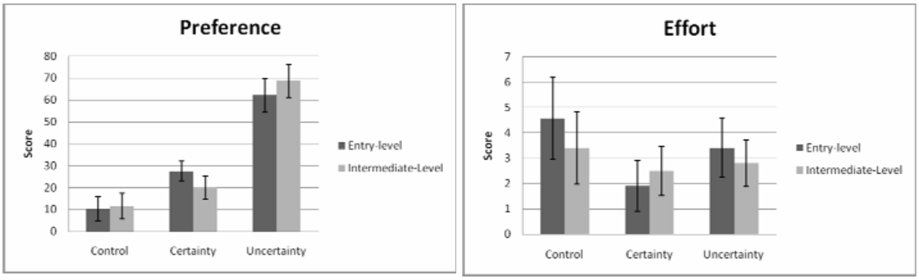


Fig. 2. Designers' preferences for the various design comparison methods, and the effort of using each method

Overall, both experienced and entry-level designers preferred the "uncertainty" DSS. Subjects' comments after the study confirmed this finding. They expressed that the "uncertainty" better reflected the uncertain nature of the design tasks. One subject stated, "it (the "uncertainty" DSS) adds another dimension of what I can do" and "it is intuitive to draw uncertainty as a range."

4.1 Perceived Information Sufficiency and Decision Confidence

During the study, we found that in some cases there were overlapping values between the set of alternatives considered, while in others cases there were not. Our expectation was that whenever there was overlap between the top alternatives, users of the "uncertainty" DSS would perceive less information sufficiency, and lower confidence in a decision that had to be made under such circumstances. Our results showed that this was true. There was a significant main effect of *decision making method* on *perceived information sufficiency* ($F_{(2,40)}$ = 4.158, p = 0.023) and *decision confidence*($F_{(2,40)}$ = 3.307, p = 0.047). When there was an overlap between top alternatives, entry-level designers who used the "uncertainty" DSS perceived less information sufficiency ($F_{(1,10)}$ = 8.095, p = 0.017), and expressed less confidence

($F_{(1,10)}$= 3.447, p = 0.093) in their decisions, as shown in Figure 3. In the control, and when using the "certainty" DSS, users did not perceive and difference in information sufficiency, regardless of whether the alternatives overlapped or not, nor did their decision confidence change.

We did not find significant effects for the intermediate-level designers, but the trends were in the same direction.

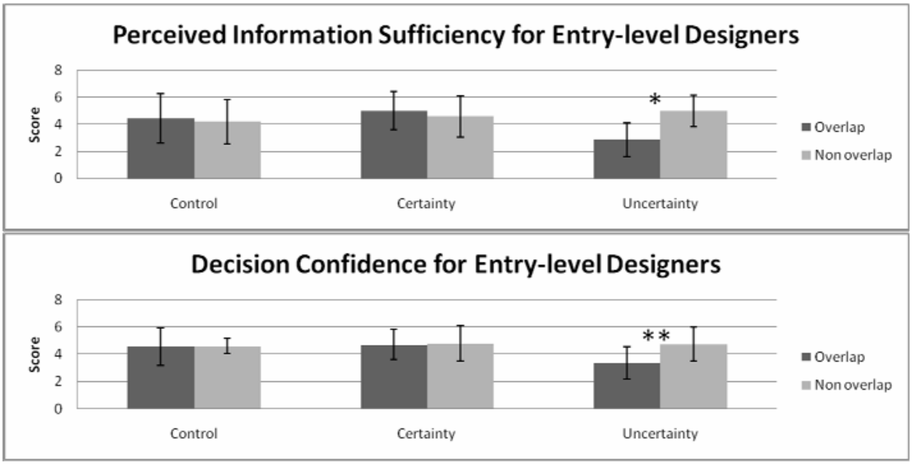


Fig. 3. Entry-level designers found the "uncertainty" tool to help them identify situations in which they lacked sufficient information to make a decision. *Significant at the 0.05 level. ** Significant at the 0.10 level.

We feel that this demonstrates that the visualization of uncertainty did help users, particularly entry-level designers to correctly identify when they lacked sufficient information to make a decision. These findings are in line with the research reported by Cole [11] and Hoffman et al. [12] who found that visualization of uncertainty information helps convey the sense of uncertainty. Also, Nadav-Greenberg [13] found that decision makers are able to recognize the value of uncertainty information, and evaluate the information to have a more realistic understanding of the situation.

4.2 Plans to Seek Additional Information

While it is important to recognize when one does not have enough information to make an informed decision, it is also important to take action to get more information, and to get the right information. We also analyzed to what extent users realized what specific information they should get to most reduce the overlap between alternatives, and whether they followed through on these intentions a week later. The criteria that contribute most to the overall uncertainty are those that have a high importance weight and have great uncertainty in their values. A very uncertain parameter which is not very important does not contribute greatly to the overall uncertainty for an alternative.

We wished to see whether the criteria which designers planned to investigate further correlated to the criteria that actually contributed most to the overall uncertainty. The results of correlation are shown in Figure 4. While the correlations were positive, they were not strong enough for either the entry-level (r = 0.418, n = 48) or intermediate-level designers (r = 0.551, n = 31). From this we concluded that they could use more assistance identifying what specific information would be most useful for clarifying the decision.

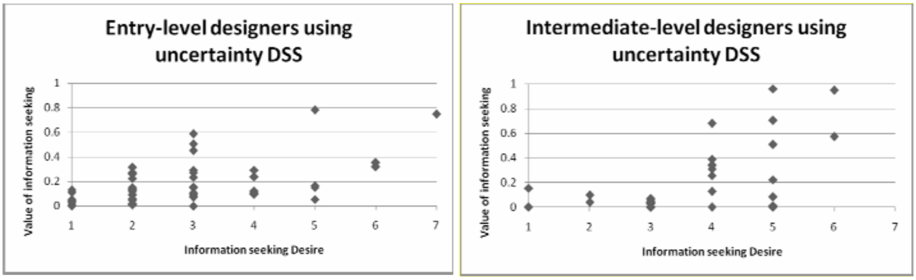


Fig. 4. Correlation results of information seeking desire and value of information seeking

One week after the experiment, we sent a follow-up survey to the participants to inquire what they did in the week following the experiment. The entry-level designers were less likely than intermediate-level designers to have followed through on their expressed plans to seek information. Much of entry-level designers' information seeking effort was spent seeking information on criteria which they had identified to be non-critical, while the intermediate-level designers' more often focused their efforts on the criteria which they had identified as important. We feel further efforts are needed to identify how to help entry-level designers carry through on appropriate information seeking plans.

5 Conclusions and Future Work

The results of this study showed that use of DSS which allows designers to express uncertainty in design parameters, and more importantly, to visualize uncertainty in the overall value of each alternative helps entry-level designers to identify when information is insufficient to allow an informed choice of the "best" from a set of alternatives. Without the visualization, designers did not perceive any difference between situations in which clarifying information was needed, and those in which it was not. This relatively simple visualization empowered designers to think more clearly about the uncertainty in the design, and its implications on their decisions. Further work is needed to identify how to help entry-level designers identify what specific additional information will help the most to inform their decision, and how to motivate them to carry through on plans to gather that information.

References

1. Newell, A.: Artificial Intelligence and the concept of mind. In: Schank, R.C., Colby, K.M. (eds.) Computer Models of Thought and Language. Freeman, San Francisco (1973)
2. Simon, H.A.: The structure of ill-structured problems. Artificial Intelligence 4, 181–201 (1973)
3. Dwarakanatha, S., Wallace, K.M.: Decision-making in engineering design: Observations from design experiments. Journal of Engineering Education 6(3), 191–206 (1995)
4. Lewis, K.E., Chen, W., Schmidt, L.C.: Decision making in engineering design. American Society of Mechanical Engineers (2006)
5. Badke-Schaub, P.: Decision making processes and leadership in product development departments. In: Proceedings of the Eighth International NDM Conference. Pacific Grove, CA (2007)
6. Tversky, A.: Preference, Belief, and Similarity: Selected Writings. MIT Press, Shafir (2003)
7. Hayes, J.R.: The Complete Problem Solver. The Franklin Institute Press, Philadelphia (1981)
8. Erev, I., Bornstein, G., Wallsten, T.W.: The negative effect of probability assessment on decision quality. Organizational Behavior and Human Decision Processes 55, 78–94 (1993)
9. Hayes, C.C., Akhavi, F.: Cost Effective Decision Aids for Complex Tasks. Journal of Usability Studies 2(4), 152–172 (2008)
10. Hayes, C.C., Akhavi, F.: Combining Naturalistic and Mathematical Decision Aids to Support Product Design. In: Proceedings of NDM 2009, the 9th International Conference on Naturalistic Decision Making, London, UK (2009)
11. Cole, W.G.: Understanding Bayesian Reasoning Via Graphical Displays. In: Proceedings of CHI 1989 Human Factors in Computing Systems, pp. 381–386. ACM Press, Austin (1989)
12. Hoffman, J.R., Wilkes, M.S., Day, F.C., Bell, D.S., Higa, J.K.: The roulette wheel: An aid to informed decision making. PLoS Medicine 3, 743–748 (2006)
13. Nadav-Greenberg, L., Joslyn, S.L.: Uncertainty Forecasts Improve Decision Making Among Nonexperts. Journal of Cognitive Engineering and Decision Making 3(3), 209–227 (2009)

Consideration of Human Factors for Prioritizing Test Cases for the Software System Test

Christoph Malz[1], Kerstin Sommer[2], Peter Göhner[1], and Birgit Vogel-Heuser[2]

[1] Institute of Industrial Automation and Software Engineering, Universität Stuttgart,
Pfaffenwaldring 47, 70569 Stuttgart, Germany
[2] Department of Information Technology in Mechanical Engineering, TU München,
Boltzmannstr. 15, 85748 Garching, Germany
{Christoph.Malz,Peter.Göhner}@ias.uni-stuttgart.de,
{sommer,vogel-heuser}@itm.tum.de

Abstract. A big challenge of software test managers is the limited test time. Especially the system test, where the whole integrated software system is tested shortly before delivery to the customer, is affected by this limitation. During the system test usually several test cycles are needed. However, a test manager cannot execute all available test cases in each test cycle due to the limited test time. He/she has to decide which test cases have to be executed in each test cycle in order to find new possible faults of the software. In this paper the Adaptive Test Management System (ATMS) based on software agents is presented which relieves the test manager from this complex manual work by using software agents for prioritizing test cases based on current information about the software system, the test cases and the human factors of the developers. The goal of the ATMS is to maximize the number of found faults in the available test time with the determined prioritization order.

Keywords: Test case prioritization, human factors, software agents.

1 Introduction

In the scope of industrial automation systems, a relatively high quality of the software system test has to be achieved by the test management in order to guarantee the required reliability of the system. One of the biggest challenges of the test manager is the limited time which he/she has for the software system test. Additionally, a software system usually has to be tested several times, due to possible software changes after each test cycle. Because of limited test time, the test manager cannot repeat all test cases in every test cycle. Somehow he/she has to prioritize the test cases in order to execute the most important ones, i.e. he/she has to find a prioritization order of the test cases that increases the probability to find more faults in the available test time.

Currently, available test management tools [1] offer support for prioritizing test cases by administrating different useful data. However, the big amount of data has to be evaluated by the test manager himself in order to find a prioritization order of the test cases. Automated prioritization techniques, as described in [2], prioritize test cases based on exact information about source code changes. However, they are not

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 303–312, 2011.

appropriate for complex software systems. Another massive drawback of the existing approaches is the fact that they neglect human factors for prioritizing the test cases. However, since the goal of the prioritization is to increase the probability to find more faults in the available test time and since the faults are made by the developers of the software system, human factors of the developers play an important role in finding the right prioritization order.

A review of the relevant research literature shows that psychological studies as to human factors influencing the performance of programmers have been accomplished since the 1950s. But from the 1990s the focus of interest shifted towards graphical user interfaces (GUIs) resulting in notedly less publications about programming abilities.

Unfortunately, the correlations between the developed programming tests and the job performance of programmers were often quite low. Curtis [3] ascribes that to researchers' deficient understanding of the cognitive requirements of programming, which often resulted in mixing different levels of analysis (e.g. using a test of general programming ability to predict debugging skills). Another reason may be the "well-known failure of a manager's performance ratings to accurately measure individual performance" [3]. Nevertheless, there are undeniably huge differences in performance among programmers when confronted with the same task and there must be something to account for this phenomenon.

According to [3], [4] and [5] the following factors should be considered in order to predict individual programming performance:

- Individual knowledge about the task, the system architecture and the system interface
- Practical experience
- Intellectual aptitudes
- Mental abilities
- Cognitive style (see also [6])
- Motivational structure (especially intrinsic motivation) (see also [7])
- Personality characteristics (e.g. self-confidence) (see also [8])
- Behavioral characteristics

However, only few authors make concrete statements about the degree to which one of these factors affect the performance of the programmer and how they are interrelated. This may be due to the usually very restricted set of affecting variables studied in the same set of data.

Considering previous results [5] as well as ease of measurement, it seems that past practical programming experience is the most promising of the assumed influencing factors.

This and other human factors are considered in the Adaptive Test Management System (ATMS) [9] which is presented in this paper. It automatically prioritizes available test cases for each test cycle with the goal to get a prioritization order of the test cases which increases the probability to find more faults in the available test time. In our approach the ATMS is realized using autonomous software agents [10]. Thereby, each test case and each software module of the software system is represented by a software agent. The software agents prioritize the test cases from their local perspective and in cooperation with the other agents, which reduces the complexity of the prioritization problem.

In Section 2 the idea of the ATMS is described in more detail. Section 3 discusses the accounting for human factors for test case prioritization. In Section 4 we present our agent-based prioritization concept. Section 5 concludes this paper.

2 Adaptive Test Management System

The task of the Adaptive Test Management System (ATMS) is the prioritization of test cases for the system test. The goal is to get a prioritization order that allows finding more faults in the available test time.

The ATMS prioritizes the available test cases for each test cycle considering changing information for each test cycle. This is why the system is called "adaptive". Due to the prioritization values the test manager decides which test cases he/she executes in the current test cycle. The scheme of the ATMS is depicted in Fig. 1.

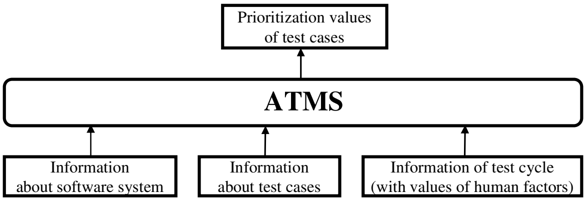

Prioritization values of test cases
ATMS
Information about software system
Information about test cases
Information of test cycle (with values of human factors)

Fig. 1. Scheme of ATMS

The ATMS has three kinds of input. First, it gets information about the software system. The software system consists of several software modules, as shown for a simple example in Fig. 2.

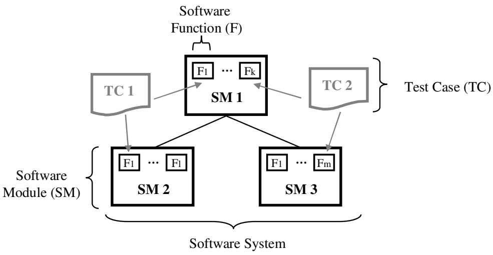


Fig. 2. Software modules, software functions and test cases

The input is a formalized architecture model out of which the ATMS derives information about the existing software modules and their functions, about the calling relations between the software modules and about other properties of the software modules.

Second, the ATMS gets information about the available test cases. This means, the test cases are already specified and can also be even implemented as automated test scripts. In detail, the ATMS gets the information about which test cases exist, information about different properties of the test cases and about the functions of a software module which a test case executes while running. As depicted in Fig. 2, a test case can execute different functions of different modules during a run. All the information about a test case is retrieved out of a test management tool and its database.

The third and last kind of information is the changing information for each test cycle. The considered information is:

- Information from the test, e.g. information about faults that have been found and their mapping to the software modules.
- Information from the development, e.g. information about software changes, i.e. which function in which software module was changed.
- Information about the values of human factors.

All this information for each test cycle is retrieved out of the databases of the test, fault and change management tools or provided by the test manager.

The outputs of the ATMS are prioritization values for all available test cases. For each test case a value between 0 and 10 is determined as prioritization value. A value of 10 implies very high priority whereas a value of 0 implies very low priority.

After describing the scheme of the ATMS with its goal, task, inputs and outputs, we discuss the account for human factors for test case prioritization in the next section.

3 Accounting for Human Factors for Test Case Prioritization

One of the biggest challenges for the prioritization of test cases using human factors – besides the selection of appropriate influencing factors – is the operationalization of these variables. It is essential that the classification of each programmer is as quick and easy as possible. Ideally, it can be integrated into the everyday work and does not require permanent or recurrent psychological testing.

Bearing this in mind, the most promising of the factors mentioned in the introduction is past practical programming experience. It has shown to have significant effects on programming performance and it can be easily quantified. At the moment it is provided to the ATMS by the test manager judging the experience of the developer with regard to the implementation work he/she has to do. According to Curtis [3], that seems to be rather unproductive. In adoption of the approach described in [5], we plan to prove an experience rating according to the qualification background of the developers (master, bachelor, technician etc.). Moreover, the years of relevant professional experience of the developers will be included. We hypothesize, that the more practical experience (gained during an apprenticeship and/or on the job) a programmer has, the fewer are the faults in the developed software system.

In addition to their practical experience, the workload of the developers implementing a software module of the software system will be taken into account. Currently, it is derived from information about software changes made by a developer, i.e. from the amount of changes, the effort for a change and the time used for implementing the changes. Thereby, the time, when the changes were made, also plays a role, e.g. were the changes made during the day, in the evening or even at the weekend. A major disadvantage of this approach is the fact that it is more a measure of task difficulty than a measure of workload. That is, it does not account for individually varying coping of the developers when confronted with the same task. In human factors research three valid approaches for measuring workload exist: Performance on a secondary task, physiological reactions (e.g. heart rate variability) and self-reported measures. Because of the mentioned importance of application ease and integration into everyday work (i.e. low task intrusion) a self-assessment of the developers on a one-dimensional rating-scale as to their felt workload seems to be the most appropriate approach. Moreover, compared to the alternatives this method does not require particular equipment, is easily accepted by the developers and produces very low costs. We assume that higher self-reported workload is associated with poorer performance of the programmer. However, one should bear in mind that workload is no static measure but is in a constant state of flux. Therefore, it has to be re-entered by the developers subsequent to the completion of every single programming task and before a new test cycle.

Measuring the other human factors that have been mentioned in the literature – i.e. intellectual aptitudes, cognitive style and motivational structure– is much more complicated. First, most of these factors identify merely the general category without any specification. However, the possible measurement depends greatly on what exactly is meant by the term. Second, some factors require an elaborate testing carried out by psychologists. For that reasons, the mentioned factors will only be included in the final testing procedure after their practical impact on prioritizing test cases for the software system test has been experimentally verified.

After discussing accounting for human factors for test case prioritization, we describe the realization of the ATMS in the next section and explain the functionality of the ATMS.

4 Agent-Based Test Case Prioritization

As described in the first section, the prioritization of the test cases for the system test is executed from the perspective of the test cases and software modules of the software system. Therefore, test cases and software modules are represented by software agents. These software agents realize the ATMS. The software agents together determine the priority of each test case using different information. We call this priority from now on the "global" priority. This means, this is the priority of a test case in comparison to all other test cases. For determining this global priority, the software agents representing the software modules, called the SM-agents, determine a test importance of a software module using information about human factors amongst others. The software agents representing the test cases, called TC-agents, determine local priorities of a test case. Combining test importances of the software modules and local priorities of the test cases, the global priority of a test case is calculated. These steps will be described in more detail in the next sections.

4.1 Perspective of the SM-Agent: Determination of the Test Importance of a Software Module

A SM-agent is generated for each software module specified in the architecture model. The goal of the SM-agent is to increase the number of found faults in the available test time. Therefore, its responsibility is to determine the test importance of the software module. This test importance indicates the probability with which faults can be found in the system by testing this software module. For determining the test importance the SM-agent retrieves different information from its environment, i.e. from other SM-agents, from the architecture model, from the available databases and from the test manager.

In the current configuration of the ATMS the following information is used for determining the test importance of a software module:

- Complexity of the software module: A more complex software module indicates a high probability for faults in the software module. The value for complexity has to be provided by the developers in the architecture model.
- Number of faults in the software module in the previous test cycle: Faults in a software module indicate a high probability for more faults. The value for the number of faults is retrieved out of the database of the fault management tool.
- Number of changes of the software module: A software module which was changed indicates a higher probability for faults in the software module. The number of changes of the software module is retrieved from the database of the change management tool.
- Number of changes in other modules in which functions are called: Impacts of faults due to changes in other modules can be found in related software modules. Therefore, the test importance of software modules which call changed software modules is increased. The SM-agent is informed by the other agents about changes of other software modules.
- Values of human factors which are currently considered:
  - Workload of the developer of the software module: A developer with a bigger workload is assumed to make more faults. Thus, a higher workload indicates a higher probability for faults in the software module. Currently, the workload is derived from the number of changes, the effort for a change and the time used for implementing the changes. All these values are retrieved from the database of the change management tool.
  - Experience of the developer of the software module: A developer with less experience in a certain task is assumed to make more faults. Thus, a lower experience of the developer indicates a higher probability for faults in the software module. Currently, the experience value is given by the test manager.

The SM-agents determine the test importance of the software module using this information.

Additionally, they need knowledge for evaluating the information. As shown in [11], the knowledge of the test manager is imprecise and can be made available as verbally formulated rules, e.g. “If the workload of the developer of a software module is high, the test importance of the software module is very high”. Furthermore, the determination of the test importance is not a YES/NO decision. It is based on indicators and always afflicted with uncertainty. Therefore, we decided to use fuzzy logic to specify the rules for the determination of the test importance. Fuzzy logic offers a good possibility to reproduce such human evaluation standards, thinking patterns and conclusions, as shown in [12].

Every SM-agent determines the test importance for the software module it represents autonomously by an internal fuzzy-unit, as depicted in Fig. 3.

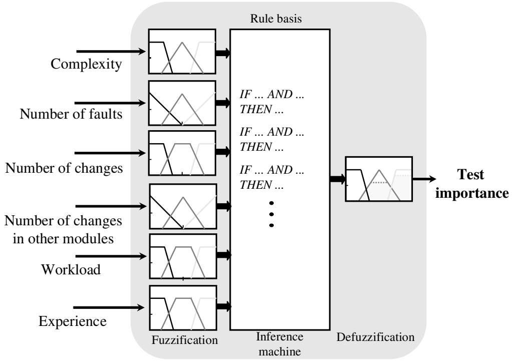


Fig. 3. Fuzzy-unit for determination of the test importance

First, the input data is fuzzified, this means crisp values are transformed into grades of membership for linguistic terms (e.g. low, middle, high) of fuzzy sets. A membership function is used to associate a grade to each linguistic term.

The fuzzified input data is then combined together by the fuzzy inference machine with rules of the type "if … and … then…" For example, rules for the workload are specified like this:

- If (workload is high) then (test importance is very high)
- If (workload is middle) then (test importance is high)
- If (workload is low) then (test importance is low)

Similarly**,** the rules for all input data are described. Since a crisp value of input data can have several grades of membership for different linguistic terms, no unwarranted precision is introduced in the prioritization.

Finally, the test importance is determined through defuzzification. The resulting test importance value has a range from 1 (very low) to 10 (very high).

After considering the perspective of the SM-agents in this section, we want to consider the perspective of the TC- agent in the next section.

4.2 Perspective of the TC-Agent: Determination of the Local Priorities of a Test Case

For each test case specified in the database of the test management tool a TC-agent is generated. The goal of the TC-agents is, as it is for the SM-agents, to increase the number of found faults in the available test time. Therefore, its first capability is to determine the local priorities of a test case with respect to each software module, which the test case executes when it runs. Thus, the local priority (LP_{xi}) is the priority of a test case "x" with respect to a software module "i" (see Fig. 4).

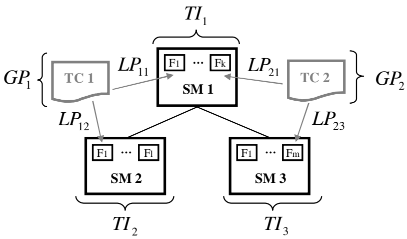


Fig. 4. Test importance, local priority and global priority

The local priority indicates the probability with which the test case can find faults in the related software module. A test case can have different local priorities with respect to different software modules. For determining the local priorities the TC-agent retrieves different information from its environment, i.e. from the available databases and from the SM-Agents.

In the standard configuration of the ATMS the following information is used for determining the local priority of a test case with respect to a software module:

- Number of found faults in the software module by the test case in average: A higher number indicates a high probability with which the test case can find faults in the related software module. The value for the number of found faults is retrieved from the database of the fault management tool.
- Number of changed functions in the software module which are executed by the test case: When a test case executes changed functions in a software module, this indicates a higher probability with which the test case can find faults in the related software module. In order to derive this information the TC-agents ask the related SM-agents if any software modules have been changed. In case of changes, the TC-agent gets information about the changed functions from the SM-Agents.

With this information each TC-agent determines the local priorities of its test case autonomously for each software module which its test case executes. As the rules for determining the local priorities have the same characteristics as for determining the test importance, we also decided to use fuzzy logic. Thus, as for the test importance, an internal fuzzy-unit is used for the determination of the local priorities in each TC-agent.

The resulting values for the local priorities have a range from 1 (very low) to 10 (very high) as for the test importance.

At this time, the software agents have determined the test importances (TI_i) of the software modules “i” and the local priorities (LP_{xi}) of the test cases “x” in parallel. The next step is to determine the global priority (GP_x) of each test case “x”, as depicted in Fig.4. This is described in the next section.

4.3 Perspective of the TC-Agent: Determination of the Global Priority of a Test Case

As mentioned above, the global priority is the priority of the test case in comparison to all other test cases. It is also determined out of the perspective the TC-agent. The

TC-agent determines the global priority based on information about the local priorities and the test importance. The TC-agent has the information about the local priorities. The information about the test importance of the software modules, which are executed by its test case, is requested by the TC-agent from the corresponding SM-agents.

The global priority is calculated by the TC-agent as the weighted average value of the local priorities:

$$GP_x = \frac{\sum_{i=1}^{n} LP_{xi} TI_i}{\sum_{i=1}^{n} TI_i} \qquad \begin{array}{c}\text{Weighting factor}\\ TI_i \geq 0 \\ \text{n = number of software modules}\end{array} \tag{1}$$

This means the local priorities are weighted by the test importances and then added together. They are divided by the sum of the test importances for scaling. In the example of Fig. 4 the global prioritiy GP_1 is calculated as follows:

$$GP_1 = \frac{(LP_{11} * TI_1) + (LP_{12} * TI_2)}{TI_1 + TI_2 + TI_3} \tag{2}$$

Considering the global priorities the test cases can be compared and ordered. The test manager can decide up to which priority value test cases should be executed in a test cycle.

5 Conclusion

This paper presented an approach for prioritizing test cases by software agents for software system test considering human factors. Software modules of the software system and test cases are represented by software agents. Using rules based on fuzzy logic the agents prioritize the test cases for each test cycle. The reached prioritization order allows finding more faults in the available test time.

A prototype of the ATMS was implemented using the agent development framework JADE and the fuzzy logic package jFuzzy Logic. The implemented prototype is connected to the databases of the test management tool Testopia and the fault and change management tool Bugzilla. The architecture model is currently provided to the ATMS in XML format. The experience of a developer is provided to the ATMS by the test manager using the GUI. Currently, we are evaluating the prototype using data from a finished software project of a company from the automotive industry in order to compare which test cases were executed without ATMS and which would be executed with ATMS. The first evaluation results are very promising. After software changes the ATMS prioritizes the test cases in a way that many test cases with very low priority can be excluded. Without ATMS almost all of such test cases were repeated after the software changes but did not find any faults. Currently, we are continuing the evaluation with more data from a bigger software project.

Furthermore, we will extend the consideration of human factors for the prioritization of test cases by selecting further human factors due to experiments and by improving the operationalization of the human factors.

References

1. Illes, T., Pohlmann, H., Roßner, T., Schlatter, A., Winter, M.: Software –Testmanagement. Heise Verlag, Hannover (2006)
2. Rothermel, G., Untch, R.H., Harrold, M.J.: Prioritizing Test Cases For Regression Testing. IEEE Transactions on Software Engineering 27(10) (2001)
3. Curtis, B.: Five Paradigms in the Psychology of Programming. In: Helander, M. (ed.) Handbook of Human Computer Interaction, Elsevier Science B.V., North-Holland (1988)
4. Carroll, J.: Mental Models in Human-Computer Interaction. In: Helander, M. (ed.) Handbook of Human Computer Interaction, Elsevier Science B.V., North-Holland (1988)
5. Friedrich, D., Vogel-Heuser, B.: Evaluating the Benefit of Modelling Notations for PLC-Programming Quality. In: Human Computer Interaction (HCI), Las Vegas(2005)
6. Bishop-Clark, C.: Cognitive Style, Personality, and Computer Programming. Computers in Human Behavior 11(2), 241–260 (1995)
7. Bergin, S., Reilly, R.: The influence of motivation and comfort-level on learning to program. In: Romero, P., Good, J., Acosta Chaparro, E., Bryant, S. (eds.) 17th Workshop of the Psychology of Programming Interest Group, Sussex (2005)
8. Bergin, S., Reilly, R.: Programming: Factors that Influence Success. In: Proceedings of the thrity-fifth SIGCSE technical symposium on Computer Science Education, St. Louis (2005)
9. Malz, C., Jazdi, N.: Agent-based test management for software system test. In: Malz, C., Jazdi, N. (eds.) IEEE International Conference on Automation, Quality, Testing, Robotics (2010)
10. Wooldridge, M., Jennings, N.R.: Intelligent Agents: Theory and Practice. Knowledge Engineering Review, vol 10(2), 115–152 (1995)
11. Xu, Z., Gao, K., Khoshgoftaar, T.M.: Application of Fuzzy Expert System In Test Case Selection For System Regression Test. In: Xu, Z., Gao, K., Khoshgoftaar, T.M. (eds.) IEEE International Conference on Information Reuse and Integration, pp. 120–125 (2005)
12. Avineri, E., Köppen, M., Dahal, K., Sunitiyoso, Y., Roy, R.: Applications of Soft Computing - Updating the State of the Art. Springer, Heidelberg (2009)

Cognitive Engineering of Automated Assembly Processes

Marcel Ph. Mayer[1], Barbara Odenthal[1], Carsten Wagels[2], Sinem Kuz[1],
Bernhard Kausch[1], and Christopher M. Schlick[1]

[1] Institute of Industrial Engineering and Ergonomics of RWTH Aachen University,
Bergdriesch 27, D-52062 Aachen, Germany
[2] Laboratory for Machine Tools and Production Engineering of RWTH Aachen University,
Steinbachstraße 19, D-52056 Aachen, Germany
m.mayer@iaw.rwth-aachen.de

Abstract. A novel approach to cognitive automation of assembly processes is introduced. An experimental assembly cell with two robots has been designed to proof the concept. The cell's numerical control – termed a cognitive control unit (CCU) – is able to simulate human information processing at a rule-based level of cognitive control on the basis of the SOAR cognitive architecture. Thus the CCU can plan assembly processes autonomously and can react to changes in assembly processes due to increasing number of products that have to be assembled in a large variety in production space as well as changing or uncertain conditions. To develop a "Humanoid-Mode" for automated assembly systems similar to the H-metaphor for automated vehicles human assembly strategies where identified in empirical investigations and formulated as production rules. When the CCU is enriched with these production rules underlying human heuristics, a significant increase of the predictability of a robot when assembling products can be achieved.

Keywords: Cognitive Automation, SOAR, Assembly, Joint Cognitive Systems.

1 Introduction

In high-wage countries many manufacturing systems are highly automated. The main aim of automation is usually to increase productivity and reduce personnel expenditures. However, it is well known that highly automated systems are very investment-intensive and often generate a non-negligible organizational overhead that is mandatory for production scheduling, numerical control programming or system maintenance, but does not directly add value to the product to be manufactured. Highly automated manufacturing systems therefore tend to be neither efficient enough for small lot production (ideally one piece) nor flexible enough to handle products to be manufactured in a large number of variants. Despite the popularity of strategies for improving manufacturing competitiveness like agile manufacturing (ZHANG & SHARIFI 2000) that consider humans with their specific knowledge, skills and abilities to be the most valuable factors of enterprises, one must conclude that especially in high-wage countries the level of automation of many production systems

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 313–321, 2011.

has already been taken very far with relatively little consideration given to the human operator.

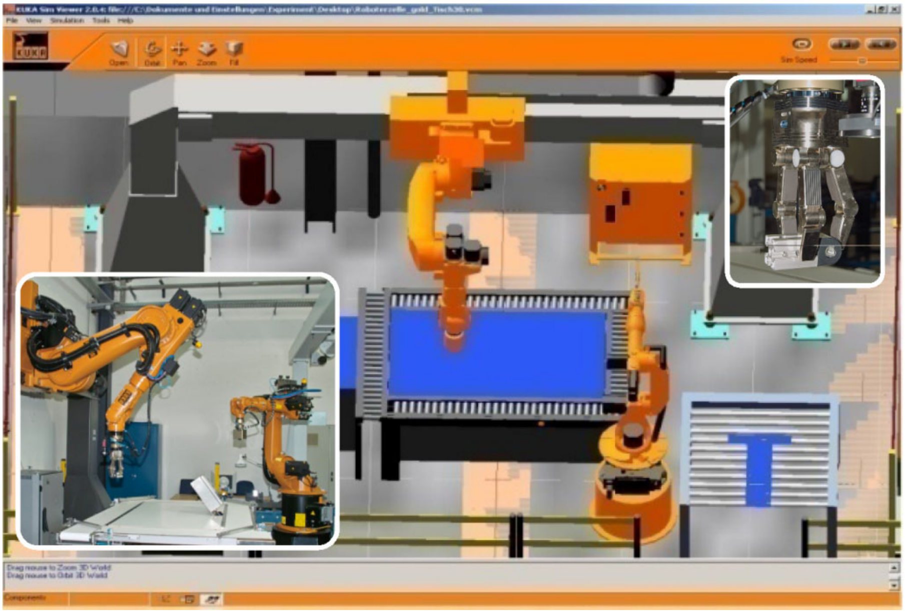
In order to achieve a sustainable competitive advantage for manufacturing companies in high-wage countries with their highly skilled workers, it is therefore not promising to further increase the planning orientation of the manufacturing systems and simultaneously improve the economies of scale. The primary goal should be to wholly resolve the so-called polylemma of production, which is analyzed in detail in KLOCKE (2009). Therefore, according to some kind of "law of diminishing returns" a naive increase in automation will likely not lead to a significant increase in productivity but can also have adverse effects. According to KINKEL et al. (2008) the amount of process errors is on average significantly reduced by automation but the severity of potential consequences of a single error increases disproportionately. These "Ironies of Automation" (BAINBRIDGE 1987) which were identified by Lisanne Bainbridge as early as 1987 can be considered as a vicious circle (ONKEN & SCHULTE 2010), where a function that was allocated to a human operator due to poor human reliability is automated. This automation results in higher function complexity, finally increasing the demands on the human operator for planning, teaching and monitoring, and hence leading to a more error-prone system. To reduce these potential errors one could again extend automation and reinforce the vicious circle. During the first turn it is quite likely that the overall performance of an automated system will increase, but the potential risk taken is often ignored or severely underestimated. Additional turns usually deteriorate performance and lead to poor solutions.

2 Cognitive Control Unit

One of today's challenges in production is the increasing complexity of assembly processes due to an increasing number of products that have to be assembled in a large variety in production space (WIENDAHL et al. 2007). Whereas in conventional automation each additional product or variant increases non-value adding processes, cognitively automated assembly cells are able to (semi-)autonomously plan and execute given tasks on the basis of a digital model of the product to be assembled. Therefore, these systems allow for flexible, cost effective and safe assembly.

The novel concept of cognitive automation by means of simulation of human cognition within the technical system aims at breaking the cited vicious circle. Based on artificial cognition, technical systems shall not only be able to (semi-) autonomously perform process planning, adapt to changing manufacturing environments and be able to learn from experience to a certain degree but also to simulate goal-directed human behavior and therefore significantly increase the conformity with operator expectations. Clearly, knowledge-based behavior in the true sense of RASMUSSEN (1986) (and also skill-based behavior to a non-negligible extent) cannot be modeled and simulated and therefore the experienced machining operator plays a key architectural role as a competent problem solver in unstable and non-predictable situations.

In order to study human-machine interaction in cognitively automated manufacturing systems, an experimental assembly cell (see Fig. 1) was designed and a manufacturing scenario was developed by KEMPF et al. (2008). The scenario is as follows: An engineer has designed a mechanical part of medium complexity by composing it e.g. with a CAD-system containing any number of subparts. The task for the assembly cell's cognitive control unit (CCU) is to autonomously develop and execute a time and energy efficient assembly sequence on the basis of the CAD model using the given technical resources in terms of robots, manipulators, changing devices, supplied subparts etc.


KUKA Sim Viewer 2.0.4: file:///C:\Dokumente und Einstellungen\Experiment\Desktop\Roboterzelle_gold_Tisch30.vcm
File
View
Simulation
Tools
Help
KUKA
Open
Orbit
Pan
Zoom
Fill
Sim Speed
Components

Drag mouse to Zoom 3D World
Drag mouse to Orbit 3D World

Fig. 1. Design of the prototypical assembly cell

MAYER et al. (2009) presented a CCU using the cognitive architecture SOAR (LEHMAN et al. 2006) to simulate cognitive functions. As outlined by MAYER et al. (2008) it is crucial for the human operator to understand the plan of the CCU to supervise the robotic assembly cell. Therefore, the question arises how the symbolic representation of the knowledge base of the CCU must be designed to ensure the conformity with the operator's expectations. Proprietary programming languages that are used in conventional automation have to be learned domain specific and do not necessarily match the mental model of the human operator. In terms of a human centered description for matching the process knowledge to the mental model, one promising approach in this particular manufacturing scenario is the use of motion descriptors, since motions are familiar to the human operator from manually performed assembly tasks (GAZZOLA et al. 2007). These motions are also easy to anticipate in human-robot interaction. Hence, already established methods or taxonomies for manual process planning should be used. In production systems it is best practice to break down complex handling tasks into fundamental motion elements. To do so the very popular MTM system as a library of fundamental movements was chosen to define the motion descriptors that can be used by the CCU to plan and execute the robotic assembly process (MAYER et al. 2008).

To implement the system, the cognitive process (CP) method as introduced by PUTZER (2004) was used. The system called SOAR-MTM was successfully evaluated in a simulation environment. In fact, all simulation runs were performed without failure and within in the expected amount of simulation cycles.

However, the system lacks of conformity with operator expectations. Regarding only one distinct pick and place operation, the expected sequence of operators – namely REACH, GRASP, MOVE, POSITION and RELEASE – was observed. The sequence of parts positioned one after another was explainable posteriori but not predictable a priori. Hence, it was not expected by the operator. Comparing the knowledge base of the CCU with the Schema Model of MARSHALL (2008), MAYER et al (2009-HCI) identified a lack of elaboration knowledge. In other words, engineering methods like the CP method (ONKEN & SCHULTE 2010, PUTZER 2004) focus more on technical aspects of cognitive systems.

When developing joint cognitive systems that have to conform to operator's expectation, it is important to acquire additional knowledge about human behavior in terms of rules and heuristics being used in manual assembly.

3 Design for Human-Machine Compatibility

In order to be able to use the full potential of cognitive automation, one has to expand the focus from a solely technical system to joint cognitive systems (HOLLNAGEL & WOODS 2005, NORROS & SALO 2009). In these systems both the human operator and the cognitive technical system cooperate effectively on different levels of cognitive control to achieve a maximum of human-machine compatibility.

To acquire additional knowledge about human behavior in terms of rules and heuristics being used in manual assembly, three fundamental assembly strategies on the basis of experimental trials with a total of 16 German subjects (basic study) could be identified and validated by a second independent experiment with a total of 25 German subjects (validation study 1; MAYER et al 2010):

- humans begin an assembly at edge positions
- humans prefer to build in the vicinity of neighboring objects
- humans prefer to assemble in layers.

3.1 Validation

The aforementioned study was conducted solely with German subjects. To account for possible intercultural influences, a second validation study was carried out with Chinese subjects (validation study 2). 11 female and 14 male subjects participated in this validation study. The average age of the 25 subjects was 22.84 years (SD: 2.03). All of the subjects had a general qualification for university entrance. All of the subjects were either still at university or had already finished their university studies. None of them normally performed manual assembly tasks in their daily work.

The assembly tasks that were given to the subjects included the assembly of 10 identical four-layer pyramids of 30 identical bricks, so that despite the laboratory conditions, a training state could be reached that would be comparable to small-series production. A detailed analysis of the acquired assembly time data can be found in Jeske et al. (2010). With respect to the boundary conditions of the robotized assembly cell, the subjects had to process the tasks under the following constraints – (1) one-handed assembly, (2) not assembling in subgroups, (3) not grasping more than one brick at a time, and (4) assembling the object on a defined working area.

After the subjects' personal data were collected, the subjects were given a written description of the assembly task. A technical drawing of the object to be assembled (azimuth of 45° and elevation of 20°) was presented on a table-mounted display. After reading the description of the assembly task, the subjects had to conduct ten trials, each of which started by double-clicking a button to start the time measurement. Then the manual assembly took place. After finishing the assembly, the subject had to double-click again to indicate the finish.

3.2 Hypothesis

If the empirically identified and validated rules hold true for the Chinese validation study, at least equal relative frequency of assembly sequences confirming the i^{th} rule (i=1,...,3) should be found in the data (see MAYER et al. 2010). On the basis of this assumption, the following hypothesis can be formulated for statistical review:

- H_i: The relative frequency of applying rule i in validation study 2 ($f_{2\_rule\ i}$) is higher than in the data collected in the basic study ($f_{basic\ rule\ i}$).
  H_{0i}: $f_{2\_rule\ i} = f_{basic\_rule\ i}$

To verify the null hypotheses, the χ^2-goodness-of-fit test was used on a significance level of $\alpha = 0.05$ due to the nominal data.

3.3 Results

The results of the χ^2-fit test are shown in Table 1. Concerning $H_{01,}$ the requirements for the χ^2-test are not met based on the observed distribution of the basic study. However, only 2.8% of the blocks were placed on internal positions, i.e. 97.2% of the blocks (100% in the basic study) were placed on edge positions, meaning that the rule can be empirically confirmed for the Chinese subjects.

According to Table 1, H_{02} cannot be rejected. Due to the very small effect size, H_{02} can be accepted. It can be said that the rule is adhered to as observed in the basic study.

Finally, according to Table 1, H_{03} must be rejected. The observed relative frequency is 86.4% (expected value from the basic study: 81.25%). The rule is therefore more closely adhered to than observed in the basic study.

Table 1. Results of the χ^2-fit test

	expected	*observed*	*p*	*X^2*	*effect size (w)*
H_{01}	1	0,972	-/-	-/-	-/-
H_{02}	0,9375	0,940	0,8703	0,0267	0,0103
H_{03}	0,8125	0,864	0,0370	4,3524	0,1319

3.4 Influence of the Rules on Prediction Accuracy of Human Behavior

To assess the predictive accuracy of human behavior, the results of the cognitive simulation will be compared to the data acquired in the empirical validation study as described in the previous section. The reference simulation model for a comparative simulation study was the basic SOAR-MTM model (MAYER et al. 2010), containing only the rules based on the fundamental motions of the MTM-1 taxonomy as well as the rules necessary to describe the assembly objects. Further, seven additional simulation models were developed. Each additional simulation model was based on the reference model but was enriched by one of the identified rules or combinations of those. An overview of the analyzed simulation models concerning the covered rule-sets is shown in Table 2.

Table 2. Overview of the compared simulation models (rule 1: edge positions; rule 2: neighborhood condition; rule 3: layer design

	MTM-1 rules	*rule 1*	*rule 2*	*rule 3*
Model 1	X			
Model 2	X	X		
Model 3	X		X	
Model 4	X			X
Model 5	X	X	X	
Model 6	X	X		X
Model 7	X		X	X
Model 8	X	X	X	X

On the basis of the criteria introduced by LANGLEY et al. (2009) for evaluating cognitive architectures, one dependent variable was formulated to assess the developed cognitive simulation models: the goodness of prediction of human assembly behavior. The goodness of prediction of a simulation model under study is defined as the probability of a given brick being correctly positioned by the simulation model during the simulated assembly sequence. The overall goodness of prediction of the simulation model is calculated on the basis of the logarithmic conditional probability (*LCP*):

$$LCP = \sum_{i=2}^{30} \log_{10} p(x_i \mid x_{i-1}) \cdot p(x_1) \tag{1}$$

For statistical analysis MATLAB R2010a was used. A Kruskal-Wallis analysis was performed to test against differences of the *LCP*-values of different simulation models (α=0.05). A significant effect was found (p=0.00). To further determine which pairs are significantly different, a multiple comparison test was performed, using a Bonferroni adjustment to compensate for multiple comparisons. Fig. 2 shows a boxplot of the *LCP*-values of the differing simulation models under study.

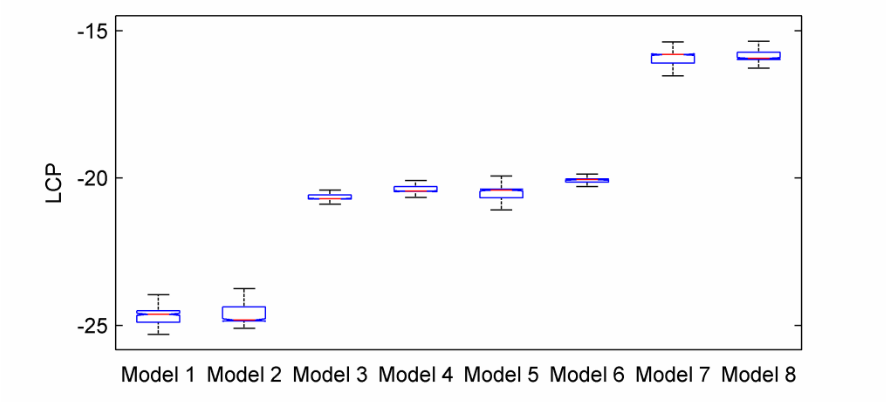


Fig. 2. Boxplot of the LCP-values of the simulation models

When comparing the simulation results, the simulation models can be assigned to three groups with significantly different predictive power. Model 1 and model 2 represent the first group with the poorest predictive power. The simulation models 3, 4, 5 and 6 are in the second group, which has medium predictive power, and model 7 and 8 belong to the third group, which has the highest predictive power.

The highest level of compatibility regarding the empirically observed human assembly operations and thus the highest prediction accuracy occurs when all rules are combined. The known overfitting-effects within research decrease the generalizability with increasing prediction accuracy. Nevertheless, 80.8% of the assembly sequences of the Chinese subjects can be simulated by this rule set. The same effect regarding predictability could be observed for the German subjects. However, MAYER et al. (2010) reported 91.6% of those assembly sequences to be covered by the same simulation model.

These result clearly shows that minor extensions to the knowledge base of a CCU can lead to a significant increase in the conformity of the assembly robot behavior with operator expectations.

4 Summary and Outlook

Especially in highly automated manufacturing systems that shall produce products in almost any variety in product space, an increase in conventional automation will not necessarily lead to a significant increase in productivity. Therefore, novel concepts towards proactive, agile and versatile manufacturing systems have to be developed. Cognitive automation is a promising approach to improve proactivity and agility. In cognitively automated systems, the experienced machining operator plays a key architectural role as a solver for complex planning and diagnosis problems. Moreover, he/she is supported by cognitive simulation models which can solve algorithmic problems on a rule-based level of cognitive control quickly, efficiently and reliably and take over dull and dangerous tasks.

To develop a “Humanoid-Mode” for automated assembly systems similar to the H-metaphor for automated vehicles (FLEMISCH et al. 2003) identified human assembly strategies where formulated as production rules. When the introduced reasoning component is enriched with these production rules underlying human heuristics, a significant increase of the predictability of the robot when assembling the products can be achieved for both Chinese and German subjects.

For future investigations our hypothesis is as follows: If the knowledge base is enriched by human heuristics the system can be better anticipated by the human operator since it corresponds to the mental model of the assembly process. Hence, an increase in predictability leads to more intuitive human-robot cooperation and therefore increases safety significantly.

Acknowledgement. The authors would like to thank the German Research Foundation DFG for the kind support within the Cluster of Excellence "Integrative Production Technology for High-Wage Countries".

References

1. Bainbridge, L.: Ironies of Automation. In: Rasmussen, J., Duncan, K., Leplat, J. (eds.) New Technology and Human Error, Wiley, Chichester (1987)
2. Flemisch, F.O., Adams, C.A., Conway, S.R., Goodrich, K.H., Palmer, M.T., Schutte, P.C.: The H-Metaphor as a Guideline for Vehicle Automation and Interaction. NASA/TM—2003-212672 (2003)
3. Gazzola, V., Rizzolatti, G., Wicker, B., Keysers, C.: The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. NeuroImage 35, 1674–1684 (2007)
4. Hollnagel, E., Woods, D.D.: Joint Cognitive Systems: Foundations of Cognitive Systems Engineering. Taylor & Francis Group, Boca Raton (2005)
5. Kempf, T., Herfs, W., Brecher, C.: Cognitive Control Technology for a Self-Optimizing Robot Based Assembly Cell. In: Proceedings of the ASME 2008 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, America Society of Mechanical Engineers, U.S (2008)
6. Kinkel, S., Friedwald, M., Hüsing, B., Lay, G., Lindner, R.: Arbeiten in der Zukunft, Strukturen und Trends der Industriearbeit. Studien des Büros für Technikfolgen-Abschätzung bei Deutschen Bundestag – 27. edition sigma, Berlin (2008) (in German)
7. Klocke, F.: Production Technology in High-Wage Countries – From Ideas of Today to Products of Tomorrow. In: Schlick, C.M. (ed.) Industrial Engineering and Ergonomics in Engineering Design, Manufacturing and Service – Trends, Visions and Perspectives. Springer, Berlin (2009)
8. Langley, P., Laired, J.E., Rogers, S.: Cognitive Architectures: Research Issues and Challenges. Cognitive Systems Research 10(2), 141–160 (2009)
9. Lehman, J., Laird, J., Rosenbloom, P.: A gentle introduction to soar, an architecture for human cognition: 2006 update (2006), `http://ai.eecs.umich.edu/soar/sitemaker/docs/misc/GentleIntroduction-2006.pdf` (Retrieved May 17, 2010)
10. Marshall, S.P.: Cognitive Models of Tactical Decision Making. In: Karowski, W., Salvendy, G. (eds.) Proceedings of the 2nd International Conference on Applied Human Factors and Ergonomic (AHFE) Las Vegas, Nevada, USA, July 14–17 (2008)

11. Mayer, M.P., Odenthal, B., Faber, M., Kabuß, W., Jochems, N., Schlick, C.M.: Cognitive Engineering for Self-Optimizing Assembly Systems. In: Karwowski, W., Salvendy, G. (eds.) Advances in Human Factors, Ergonomics, and Safety in Manufacturing and Service Industries. CRC Press, USA (2010)
12. Mayer, M.P., Odenthal, B., Faber, M., Kabuß, W., Kausch, B., Schlick, C.M.: Simulation of Human Cognition in Self-Optimizing Assembly Systems. In: Proceedings of 17th World Congress on Ergonomics IEA 2009, Beijing (2009)
13. Mayer, M.P., Odenthal, B., Grandt, M., Schlick, C.M.: Task-Oriented Process Planning for Cognitive Production Systems using MTM. In: Karowski, W., Salvendy, G. (eds.) Proceedings of the 2nd International Conference on Applied Human Factors and Ergonomic (AHFE). USA Publishing, USA (2008)
14. Norros, L., Salo, L.: Design of joint systems: a theoretical challenge for cognitive system engineering. Cognition, Technology and Work 11, 43–56 (2009)
15. Onken, R., Schulte, A.: System-ergonomic design of cognitive automation. Studies in Computational Intelligence. Springer, Berlin (2010)
16. Putzer, H.J.: Ein uniformer Architekturansatz für kognitive Systeme und seine Umsetzung in ein operatives Framework. Köster, Berlin (2004) (in German)
17. Rasmussen, J.: Information Processing and Human-Machine Interaction. An Approach to Cognitive Engineering. North-Holland, New York (1986)
18. Wiendahl, H.P., ElMaraghy, H.A., Nyhuis, P., Zäh, M.F., Wiendahl, H.H., Duffie, N., Brieke, M.: Changeable Manufacturing. Classification, Design and Operation. Annals of the CIRP 56(2), 783–809 (2007)
19. Zhang, Z., Sharifi, H.: A methodology for achieving agility in manufacturing organizations. Int. J. Oper. Prod. Manage. 20(4), 496–512 (2000)

Delegation to Automation: Performance and Implications in Non-optimal Situations

Christopher A. Miller[1], Tyler H. Shaw[2], Joshua D. Hamell[1], Adam Emfield[2], David J. Musliner[1], Ewart de Visser[2], and Raja Parasurman[2]

[1] Smart Information Flow Technologies, 211 First St. N. #300
Minneapolis, MN USA 55401
{cmiller,jhamell,musliner}@sift.net
[2] Human Factors and Applied Cognition Program, George Mason University,
4400 University Dr MS3F5, Fairfax, VA USA 22030
{tshaw4,aemfield,edevisse,rparasur}@gmu.edu

Abstract. We have previously advocated *adaptable* interaction with automation through approaches derived from human-human delegation and using the metaphor of a sports team's "playbook". In work sponsored by the U.S. Army's Aeroflightdynamics Directorate (AFDD), we have been studying the effects of play-based delegation on human-machine system performance. Of particular interest is performance with plays in "non-optimal play environments" (NOPE) where no, or only poorly fitting, plays exist to achieve needed behaviors. Plays have been shown to offer benefits in situations for which they are customized, but more interesting is whether complacency, expectation, loss of training, and automation bias might affect performance when plays do not perfectly fit. We provide a taxonomy of NOPE conditions and report on the exploration of some of these conditions in a series of three experiments performed to date.

Keywords: adaptive/adaptable automation, playbook, delegation, automation complacency, automation bias, mixed initiative automation.

1 Introduction

We have previously advocated *adaptable* automation through flexible delegation and, the metaphor of a sports team's "playbook" (e.g. [1]). In adaptable automation, the human initiates behaviors by "delegating" desired goals, plans, constraints or stipulations at flexible levels of specificity, which automation is then responsible for executing. By contrast, more traditional *adaptive* automation approaches leave decisions about when and how to adapt to the automation. Playbook®, SIFT's approach to adaptable automation for uninhabited vehicles (UVs) allows humans to "call a play" which an automated planning and execution control system understands. The planning system then expands that play to an executable level and manages its execution, replanning as necessary within the constraints imposed in the initial play calling.

Plays are not scripts, but are templates of goals and partial plans that must be instantiated for existing circumstances when called. Plays can be, and in our work have been, represented by hierarchical task networks which embody alternate methods of

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 322–331, 2011.

performing the play. Plays can be called at a high level, in which case the operator delegates authority over all decisions which must be made about alternate subtask methods and resource usage decisions (within the "space" defined by the play definition itself) to the automation. Alternatively, the operator can "dive down" into the hierarchical structure of the play to offer increasingly specific "instructions" (constraints and stipulations) about exactly how a given instance of the play must be performed.

Previous work [2, 3, 9] has shown that flexible play-based delegation approaches provide payoffs in terms of human-machine performance across a variety of context conditions. In recent work sponsored by the U.S. Army's Aeroflightdynamics Directorate (AFDD), we have been using a multiple Unmanned Aerial Vehicle (UAV) simulation environment called MUSIM (for "Multi-UAV Simulation") to study the effects of play usage. Of particular interest has been "non-optimal play environment" (NOPE) conditions where the plays which exist provide no good fit for the circumstances. We might expect plays to offer benefits in situations for which they are customized since they offer a streamlined means of activating automation. More interesting, though, is whether complacency, expectation, loss of training, and automation bias might affect performance when plays do not fit. In this paper, we will define our use of plays, provide a taxonomy of NOPE conditions and report on the exploration of some of these conditions in a series of experiments performed to date.

2 Plays, Playbook® and Play Calling

A *play* (whether defined for automation or humans in teams) bounds a "space" of behaviors which are agreed to fall under the label of the play name. The behavioral space can be thought of as a hierarchical decomposition of alternate tasks, as in task analysis [4] and hierarchical task network planning [5]. The top level of this hierarchy represents a goal to perform the play, with various sub-goals that decompose the parent into alternate methods of accomplishment. Note, that the space does not include *all* possible behaviors the system can perform. Instead, only certain behaviors in certain combinations are agreed be exemplars of the play. For example, a "Hail Mary Pass" is a play (cf. Fig. 1) in American football in which many receivers run far downfield and the quarterback attempts to throw the ball to one. This play definition supports a wide variety of specific methods (e.g., exactly how many players run downfield, what patterns they run, when the ball is thrown, etc.) but some behavioral combinations fall outside the play definition. For example, a "Hail Mary Pass" in which zero or one player runs downfield to receive is non-sensical by definition.

There are three other important attributes of plays to be noted. First, plays can generally not be exhaustively defined in advance in a changing and incompletely knowable world. A quarterback will rarely specify a priori who he will throw the ball to, since that will be a function of who is least well defended. Second, plays demand that some autonomy be delegated to intelligent subordinates if any workload reduction and effective use of diverse skills is to be obtained. Third, the hierarchical play structure—especially its capturing of alternate methods to satisfy the play—provides a framework for conversation about exactly how the coach or quarterback intends an instance of the play to be performed—for example, stipulating how many and which receivers should go downfield and what patterns they should run to avoid confusion.

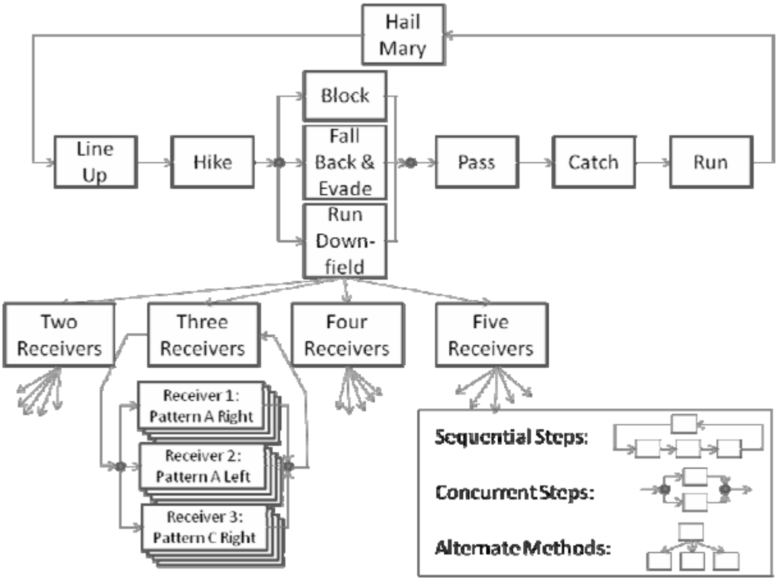


Fig. 1. Example task decomposition of a "Hail Mary" play-- showing sequential and concurrent dependencies as well as alternative methodologies

Play calling is a delegation method in which the supervisor's intent is expressed in a predefined play vocabulary. A "play" provides a label (the play's name) by which very complex behaviors can be activated quickly by a well-trained team, as well as a structure for further discussing and refining the instructed behavior. As described in prior work [1], play calling is a means of giving efficient flexibility to a supervisor over how much time s/he wants to spend in explicitly declaring intent vs. relying on subordinates' intelligence to achieve desired goals. But plays are always overlaid on a range of behaviors that is larger than that captured in the plays themselves. Players can do things, and do them in combinations, that are not be captured in the play set.

We have been developing human-automation interaction systems that provide delegation and play calling capabilities to human supervisors in interaction with smart subsystems—UAVs in most of our work. Our efforts have revolved around a core architecture we call *Playbook*® by analogy to the set of plays a sports team uses. Playbook is described elsewhere (e.g., [1]) but we will provide a brief description here.

In all Playbook architectures, the user communicates with automated systems through Playbook using a user interface (UI) configured to the domain and context of use. User instructions are interpreted by an Analysis and Planning Component which is a planning system. The UI is built around, and the Planning Component is designed to understand, a Shared Task Model which is a hierarchical decomposition of methods/plays defined. In current versions, the planner is built around a Hierarchical Task Network planner, SHOP2 [5]. The supervisor/user calls a play with whatever additional stipulations are desired via the UI. The planner then attempts to develop a plan (perhaps with existing, special purpose planners for routes, sensor coverage, etc.) which accomplishes the play in the current circumstances. If this is impossible, the planner reports it and may begin a dialogue about what is feasible. Once a plan for executing the play is agreed upon, Playbook manages the execution, adapting it within the constraints of the play called. In current versions, execution monitoring is performed by a modified version of the OpenPRS procedure-based execution management system [6]. Playbook's executive sends ongoing commands to the real or simulated control algorithms of the UAV(s) it manages, and receives updates from them to iterate through this plan management process.

3 Non-Optimal Play Environments

Plays achieve their efficiency by compiling a set of behaviors from among all those possible and assigning an easily-accessed label to them. If every possible combination had a label assigned, determining the correct one would be inefficient for both supervisor and subordinates. Thus, the play set will need to be limited for most domains. In fact, a survey of current web forums discussing the number of plays in a football team's playbook turns up answers ranging from 10-12 for a junior team to around 100 for a professional team (depending on the manner of counting variations).

Plays will typically be defined for useful behavior which either recurs frequently or which, though rare, is anticipatable and will need efficient communication and coordinated. As such, plays will necessarily capture and label some combinations of behaviors at the expense of others. A well-designed play set will provide efficiency by making critical and/or repeatedly needed complex behaviors rapidly accessible, but it will nevertheless leave some less common, less anticipated contexts less well covered. That is, available plays will be "optimal" for some contexts in that they will be easiest to command, most readily understood and provide the most effective and accurate behavior from the subordinates. But the set will inevitably be "non-optimal" for others.

In this sense, defining plays is analogous to defining automation itself. Automation makes some tasks easier, but may make rarer, less expected tasks more difficult by taking the human "out of the loop" and making him or her subject to "automation bias", complacency, and skill loss [7,8]. Might play definition may be subject to similar perils? For example, might providing a play that proves generally useful in streamlining access to a pattern of automation behaviors make it more difficult for a supervisor to access individual behaviors and combine them in a novel fashion in rare circumstance for which the set of plays provides no useful coverage?

In this research, we were particularly interested in the contrast between human performance with Playbook automation in "Non-Optimal Play Environments" (NOPE). There are various ways in which play calling can be "non-optimal":

1. *No appropriate play exists*-- there is no way for the supervisor to declare what s/he wants in a language the subordinate understands. Even here, though, especially when the intermediate levels of hierarchically decomposed plays are accessible, other plays may be useful to perform part of the desired functionality.
2. *The play is hard to command*-- The user can declare what's desired, but doing so requires excess work because the declaration "language" is not efficient. For example, excessive options must be specified for activating the desired play version (e.g., excessive tuning or stipulation requirements). This will usually result when a

default (and most easily commandable) version of a play does not fit the current need, but can be modified or further constrained to be made to fit. Thus, it is a problem of play definition rather than of UI (see below).

3. *Play communication is poor*—The play set is a good fit for the contexts and goals, but the user has difficulty communicating them. In human-human interaction, this might be due to a failure of the supervisor to enunciate clearly, or a radio channel which is full of static. In human-automation interaction, the UI itself is the problem, not the play set or reasoning about it--e.g., excessive pull down menus rather than a direct graphical or speech commanding.
4. *The play is poorly executed*– That is, the supervisor can effectively communicate intent and the subordinate(s) can understand it, but they can't perform it reliably, either due to lack of knowledge or skill or both. In these circumstances, plays are not at fault, but attempting to use them may obscure the more fundamental flaws of the subordinate agents by implying that that functionality is commandable.

In the research reported below, we have primarily focused on the first of these NOPE types—conditions in which the set of plays available, though generally useful, is lacking a play for a set of conditions that arise. The other NOPE types remain important and of interest, but investigating them must await future research.

4 NOPE Experiments and Results

Previous experimental work has shown distinct benefits in overall performance and perceived human workload for delegation-based interaction systems [2,3,9], but this might be expected if plays were optimized for the conditions in the experiments. Here, we wished to examine conditions under which plays are not optimal for at least some of the conditions which occur—and in which the operator is required to abandon play usage and instead rely on more primitive behavior commanding.

We made use of the Multi-UAV Simulation (MUSIM) testbed developed by AFDD and illustrated in Fig. 2. MUSIM simulates control and imagery from multiple UAVs (notionally, Shadows) operating simultaneously. It provides both low level (joystick) flight and sensor control and somewhat more complex, autopilot-like capabilities such as waypoint control, simple flight patterns (such as circles and racetracks), and ground target tracking. We used a hybrid version of our Playbook and the Delegation Control ("DelCon") system developed by personnel at AFDD [3] to provide play-like delegation control of these lower level behaviors in MUSIM.

In the majority of our work, the MUSIM environment has been configured to provide three UAVs with slightly different capabilities operating in an urban environment to monitor three pre-designated locations ("Named Areas of Interest" or NAIs) for civilian, military and weaponized military vehicles. UAV Alpha can only provide camera imagery; Bravo can provide imagery and can lase targets to provide shooting coordinates—but cannot itself shoot; Charlie can provide imagery and can shoot, but cannot last (and therefore must coordinate with Bravo to shoot). The user's job is to manage these three UAVs to monitor vehicles moving about the urban area, accurately discriminate military threats from civilian vehicles (by "painting" civilians to designate them as such), and then track, lase and shoot them.

Plays in Playbook and DelCon were designed to assist with these tasks and are described below. In all cases, discrimination of civilian vs. military vs. weaponized ground vehicles had to be done visually by the participant, as did steering the sensor to center the camera on a vehicle and then putting that sensor in "track mode".

1. *Monitor NAI*—one user-specified UAV is routed to and set to orbit and provide video of one user-specified NAI (from the set of three pre-designated NAIs).
2. *Lase*—Given that a vehicle is already being tracked, task a (user-specified) laser equipped UAV to lase it and pass coordinates to a missile-equipped UAV.
3. *Monitor NAIs*—Each of the three UAVs is tasked (via this single play call) to move to orbit one of the three NAIs apiece and provide video of it.
4. *Track Target and Reconfigure*—Task a user-specified UAV to continue doing what it is doing (e.g., tracking a moving ground vehicle) and task the remaining two UAVs to monitor the three NAIs (including some swapping of camera foci).
5. *Prosecute Target*—Given that a vehicle is already being tracked, task the coordination of lasing and firing at that target (after a final human authorization).

Initial studies using this set of plays in MUSIM were conducted by Fern and Shively [3] and showed significant advantages for play usage over the traditional waypoint-based flight controls representative of current operational UAVs. Fern and Shively's goal was to compare this more manual "tools" control mode vs. tools + single ship plays (1 and 2 above) vs. tools + single and multiple ship plays (1-5). Dependent variables were performance on the primary task (detecting and tracking military vehicles and prosecuting those which were weaponized) as well as on a secondary task ("painting" civilian vehicles via a designator button press). Results showed improved accuracy, reduced reaction time and reduced workload (by TLX ratings) for play usage—and these values were significant for the multi-ship plays on the secondary tasks (and

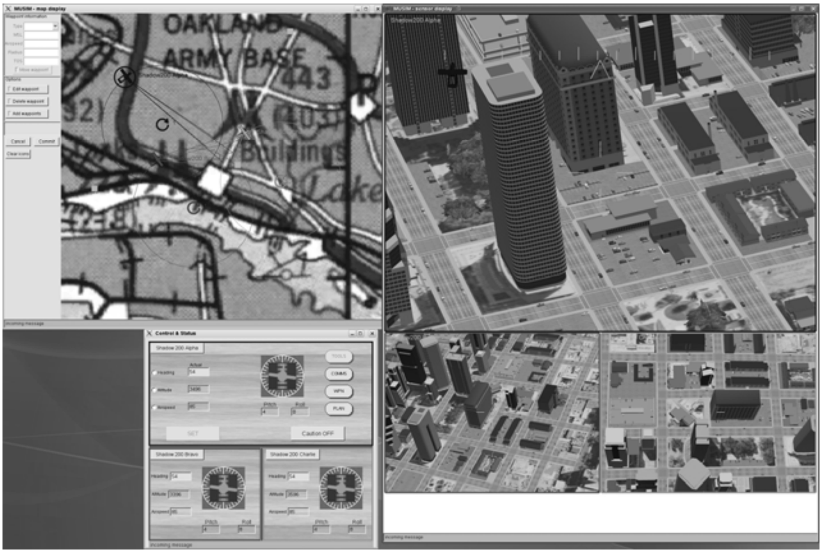


Fig. 2. The MUSIM simulation environment

in a similar direction, but not significant for primary tasks). Our own work began with a similar study varying operational tempo (through variations in the number of ground vehicles present) which showed similar results: in both high and low tempo circumstances, operators painted significantly more civilian vehicles, and tracked and prosecuted more military vehicles (non-significantly) when they had access to plays.

But again, the plays were designed to support exactly these tasks in this environment. In a subsequent pair of studies, we have included circumstances for which the set of plays is non-optimal. This has been accomplished by inserting a “popup” threat which is not in a pre-designated NAI location, and is located outside the geographic area in which the baseline tasks described above are conducted. Subjects are notified of this threat and the (high priority) need to prosecute it by means of an incoming text message. Since this popup occurs in an “undesignated” location, the play-based software the subjects use doesn’t “know” about it—which means none of the plays above can be called for it (i.e., “Monitor NAI” won’t work since it is not in a designated NAI location). Instead, subjects in all conditions have to use the “manual” level of tools to create a waypoint path for a UAV to go and find the popup threat, and then must manually lock on to it, lase it, hand off coordinates and prosecute it. Of primary interest was whether subjects who had had use to plays throughout a longer period of operating in the MUSIM environment would be at a disadvantage (either overall or specifically in dealing with the popup threat) relative to subjects who had been forced to use the more basic, waypoint-based UAV control tools throughout.

4.1 Experiment 1—Short Trials, Moderate NOPE Frequency

In our first NOPE experiment (presented in more detail in [9]), we were also interested in varying operational tempo (and therefore taskload) levels. Each of 15 participants completed 6 missions, each one including a single NOPE event. The experiment used a 2 × 3 within-subjects design with two levels of operational tempo/taskload (low, high) and three levels of control mode (Manual Tools, Single Ship Plays + Tools, and Single Ship + Multi Ship Plays + Tools). In low tempo conditions, 19 civilian vehicles were presented; in high tempo conditions, 64 civilian vehicles were presented (frequencies both higher and lower than those used in [3]). All trials lasted 10 minutes. Subjects were trained, demonstrated proficiency on the range of control modes available, and then performed their six trials in a randomized order.

Results were evaluated on the accuracy and speed of performance on the ongoing tasks (target tracking and civilian vehicles painting), as well as the ability to find and destroy the popup threat. Prosecuting weaponized vehicles occurred very infrequently and was omitted from the analysis. As with the prior studies described above, subjects generally performed better on the “normal” (i.e., non-NOPE) tasks when they had the full range of plays available. Tracking accuracy showed that participants established track on significantly (*F*(2, 28)= 4.3, p <.05) more military ground vehicles when they had the full range of plays available (M = 61.7, SE = 8.6) than when they only had manual tools (M= 46.1, SE = 6.3), with single ship plays + tools falling between those extremes (*M* = 49.4, *SE* = 6.2). For the secondary task of painting civilian vehicles, access to tools provided small advantages in accuracy and reaction time, but these did not reach significance. High operational tempo hurt performance (significantly reducing accuracy and increasing reaction time), but there was no significant interaction of tempo and control mode—indicating, that a range of plays provided benefits that were not sensitive to workload levels, at least within the ranges tested.

But performance in the NOPE events was of more interest. If an "automation complacency" effect [8] was produced by consistent use of optimal plays before the NOPE, then we expected subjects to perform worse during NOPE in trials where they had access to play-based control vs. using manual tools throughout. But this is not at all what occurred. Instead, participants were significantly ($F(2,28)= 13.48$, $p < .05$—see Table 1) faster in dealing with the popup threat when they had the full range of plays available than with manual tools alone—even though those plays were no help to them in prosecuting the popup threat. Again, the single ship plays fell between the other control conditions. Operational tempo also produced a main effect ($F(1, 14) = 11.45$, $p < .05$) such that subjects were faster to prosecute the popup in low tempo conditions, but again there was no significant interaction between tempo and control.

Thus, far from finding evidence for a complacency effect with play usage, we found the opposite. Even though plays were not helping to perform the NOPE event, having them available during the remainder of the trial helped even during NOPE. Instead of producing over-reliance on plays, or loss of ability and familiarity with manual tools, having well-fitting plays during other portions of the trial may have freed up enough cognitive workload and situation awareness capacity to allow users to "stay ahead" of the situation and better deal with the NOPE when it occurred.

4.2 Experiment 2—Longer Trials, Rare NOPEs, Sequence Variations

It might reasonably be objected that having six ten-minute trials, each containing a ~3 minute NOPE event, hardly gave time for participants to develop "complacency" in automation use. Thus, in a second experiment, we made a more rigorous attempt to induce complacency effects. Trials were 30 minutes long and contained two NOPEs. The single-ship plays control mode was dropped and we compared only manual tools vs. full range of plays (single and multi-ship). Similarly, operational tempo was not included as a variable and an intermediate tempo was used. Finally, since prior work [10] showed that when a failure is experienced (early in one's work with automation vs. after a period of reliable performance) affects trust and usage decisions, we also explored this variable by contrasting trials in which the NOPE events happened close to each other near the end of the 30 minutes vs. others in which one NOPE event happened within the first 5 minutes and the second happened at ~25 minutes into the trial. We called these the Late/Late (or L/L) vs. Early/Late (E/L) sequence conditions.

Thus, experiment 2 was a 2x2 blocked design with NOPE timing (E/L vs. L/L) being one factor and control mode (Tools vs. Plays) being the other. Each subject received two trials instead of the four required for a full between-subjects design. The different blocked combinations

Table 1. Time (in sec.) required to prosecute the NOPE event with different control modes

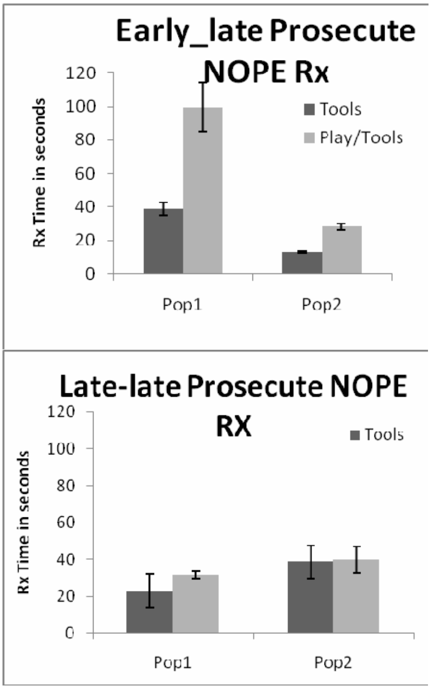
	M	SE
Manual Tools	214.22	6.19
Single Ship Plays	189.49	6.00
Single & Multi-Ship Plays	175.11	3.64

of trials were: (1) E/L + Tools, then L/L + Plays, (2) E/L + Plays, then L/L + Tools, (3) L/L + Plays, then E/L + Tools, (4) L/L + Tools, then E/L + Plays. Twenty subjects were each randomly assigned to one of the blocks.

The consistent findings from prior studies of advantages for plays on the non-NOPE tasks were largely absent here. We saw no significant effects of control mode over the full 30 minute trials. This may have been due to the added time available for participants to become familiar with the tools control mode—allowing increased competency with the more difficult manual controls to produce a ceiling effect.

There were, however, interesting findings in performance on the popup. If a complacency effect for plays exists, we expected prosecuting the popup to be slower for participants in play conditions vs. tools. This effect was expected to be larger for those who had a longer period to become complacent (those in the L/L condition).

Again, this is not what was observed (cf. Fig. 3). There was a main effect of popup sequence (*F* (1, 19) = 6.13, $p < .05$) with the second popup in each trial being prosecuted faster than the first. There was also a significant interaction of popup sequence with timing (*F*(1, 19)= 14.66, $p < .01$) such that participants were much slower to prosecute the first popup in E/L trials than in L/L trials. Fig. 3 makes it clear that this was largely due to a much slower response from subjects in the plays/tools condition. Looking only at data for E/L trials, we see that the second popup was prosecuted faster than the first (*F* (1, 16) = 13.20, $p < .01$) and that tools control alone was marginally faster (*F*(1, 16) = 3.69, *p*= .07) than control via plays + tools. The interaction between Popup position x Control mode was non-significant (*F*(1,16) = 2.23, *p* = .155) in spite of the large apparent Plays decrement for the first popup.


Early_late Prosecute NOPE Rx

Late-late Prosecute NOPE RX

Fig. 3. Reaction time results for prosecuting the NOPE in Experiment 2

But note that this is not at all what we would have expected if the use of plays in optimal conditions produced complacency and poor performance in suboptimal conditions. If that had been occurring, we should have seen greater complacency the longer subjects had to experience the optimal use of plays—in the L/L condition. Instead, participants using plays are more disrupted in the first (earliest) NOPE event they encounter. While this may be evidence of overreliance on plays, it would appear that that overreliance decreases over time, rather than increases.

5 Discussion and Conclusions

Delegation control as a model for interaction with automation (and, in particular, unmanned vehicles) is increasingly showing promise, but it is unlikely to be a panacea. After all, human-human delegation and “supervisory control” is far from perfect. We went looking for complacency and over-reliance on plays under conditions where they were not optimal, but in spite of allowing subjects up to 30 minutes to work with plays and almost 25 minutes before providing a non-optimal event, we saw no consistent evidence for such effects. The finding that plays were significantly worse in handling the *first* popup in the E/L condition in Experiment 2 is exactly the opposite of a complacency interpretation. While this may point to a need for added training to become fully comfortable with the use of plays and tools concurrently, it provides no support for the claim that plays lead to unique decrements in some circumstances.

Acknowledgments. This work was funded by a SBIR grant from the U.S. Army Aeroflightdynamics Directorate, contract # W911W6-08-C-0066. The authors would like to thank Jay Shively, Lisa Fern and Susan Flaherty for both management oversight and significant technical contributions which lay the groundwork for this work.

References

[1] Miller, C.A., Parasuraman, R.: Designing for flexible interaction between humans and automation:Delegation interfaces for supervisory control. Human Factors 49, 57–75 (2007)

[2] Parasuraman, R., Galster, S., Squire, P., Furukawa, H., Miller, C.: A flexible delegation interface enhances system performance in human supervision of multiple autonomous robots: Empirical studies with RoboFlag. IEEE Trans. Systems, Man, & Cybernetics 35, 481–493 (2005)

[3] Fern, L., Shively, R.J.: A comparison of varying levels of automation on the supervisory control of multiple UASs. In: Proc. AUVSI’s Unmanned Systems, North America, Washington, D.C. (2009)

[4] Kirwan, B., Ainsworth, L.: A Guide to Task Analysis. Taylor & Francis, London (1992)

[5] Nau, D., Au, T., Ilgami, O., Kuter, U., Muñoz, H., Murdock, W., Wu, D., Yaman, F.: Applications of SHOP and SHOP2. Technical Report, University of Maryland (2004)

[6] Ingrand, F., Georgeff, M., Rao, A.: An Architecture for Real-Time Reasoning and System Control. IEEE Expert, 34–44 (December 1992)

[7] Parasuraman, R., Riley, V.: Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors 39(2), 230–253 (1997)

[8] Parasuraman, R., Molloy, R., Singh, I.: Performance consequences of automation-induced complacency. Int. Journal of Aviation Psychology 3, 1–23 (1993)

[9] Shaw, T., Emfield, A., Garcia, A., de Visser, E., Miller, C., Parasuraman, R., Fern, L.: Evaluating the Benefits and Potential Costs of Automation Delegation for Supervisory Control of Multiple UAVs. In: 54th Meeting of the Human Factors and Ergonomics Society, pp. 1498–1502. HFES Press, Santa Monica (2010)

[10] Rovira, E., McGarry, K., Parasuraman, R.: Effects of imperfect automation on decision making in a simulated command and control task. Human Factors 49, 76–87 (2007)

Effective Shift Handover

Thomas Plocher[1], Shanqing Yin[2], Jason Laberge[1], Brian Thompson[3], and Jason Telner[1]

[1] Honeywell International, Automation and Control Solutions, Advanced Technology Laboratory, 1985 Douglas Drive, Golden Valley, MN 55422 USA
[2] Nanyang Technological University, School of Mechanical and Aerospace Engineering, 50 Nanyang Avenue, North Spine (N3), Level 2, 639798 Singapore
[3] ENGEN Refinery a division of Engen Petroleum Ltd, Process Control Department PO Box 956, Durban, 4000 South Africa
{tom.plocher,ason.laberge,Jason.telner}@honeywell.com,
Syin@pmail.ntu.edu.sg, Brian.Thompson@engenoil.com

Abstract. In the refining industry, control room and field operators document their daily activities using shift logs. These logs are supposed to be an important part of the shift handover process and are the mechanism by which activities are coordinated across shifts. Previous research identified the need for a more structured approach to shift handover. However, the value of a structured approach has never been demonstrated experimentally. We report here on an experiment sponsored by the Abnormal Situation Management Consortium conducted at the ENGEN Refinery that compared the quality of shift handovers using a structured checklist-integrated logbook to a more traditional less structured logging approach. The results showed that significant benefits to situation awareness derive from the more structured approach.

Keywords: logbook, shift work, shift handover, process control, situation awareness.

1 Introduction

Early research found that communication is one of the factors that affects abnormal situation management in the continuous process industries [1]. Information exchange between shifts is a particularly critical failure mode. Information in shift logs often is limited in usefulness by a lack of structure and poor legibility and as a result white boards, post-it notes, and change sheets are common ways of enhancing communicating and coordinating across shifts [2]. However, many of these communication mechanisms suffer from a lack of structure and permanence. Laberge, et al. [3] conducted an extensive study of requirements for effective electronic shift logs. They reviewed current logging practices in industry, user cases, and failure modes, and recommended a series of best practices for logging and shift handover. The top recommended best practice was to improve the structure of the shift handover process using structured

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 332–341, 2011.

shift logs. Hence, the Abnormal Situation Management (ASM®) Consortium (www.asmconsortium.org) funded a research project to investigate the impact that structured shift logging material has on shift handover effectiveness.

Several industrial incidents also emphasize the importance of effective logging and shift handover. On July 6, 1988, a large fire and explosion on the Piper Alpha offshore platform killed 165 and destroyed the facility. In his investigation, Cullen [4] identified several root causes and recommendations. Notably, a relief valve was removed for service and a blank had been loosely installed in its place. This information was not recorded in the control room or maintenance logs. During shift handover, the status of the pump work was discussed, but no mention was made of the relief valve work. Upon restart, the pump leaked, producing a flammable hydrocarbon cloud.

A more recent incident occurred at a BP refinery in Texas City on March 23, 2005 [5]. Fifteen people were killed and over 170 harmed as the result of a fire and explosion on the isomerization unit. The explosion occurred when a flammable vapor cloud formed following liquid overflow from the blowdown stack during operation of the raffinate splitter. The report noted several root causes, including a failure to log information and an informal and unstructured shift handover process. Both failures were contributing factors to the incident.

Collectively, the incidents and previous field research suggests that there is a need for a more efficient way to guarantee that the next shift gets the information needed for shared situation awareness. Research in other industries also suggests that better structure and organization are keys to more effective logging and shift handovers. Parke and Kanki [6] investigated the causes of documented aircraft incidents that could be traced to maintenance defects caused by a failure to communicate critical information at shift handover. A major conclusion was that face-to-face handovers are essential, but they are even better if they are supported by structured written material. They suggested a checklist of items be used to structure the shift handover. Their rationale was that written material introduces a certain redundancy in the otherwise completely verbal handover which reduces the possibility for errors in communication [7]. Parke and Kanki also point out that structuring the handover around a written checklist forces the organization to specify ahead of time the most important items of situational information for their particular operation, those items that should never be left out of the handover communication. Face-to-face shift handovers with written support have also been shown to reduce errors in aviation maintenance compared to strictly written handovers with all verbal communication filtered through a supervisor [8]. Face-to-face turnovers with written support are standard operating procedures in many high-risk domains, such as U. S. nuclear power plants [9].

In summary, there is strong direction in the literature to structure verbal face-to-face shift handovers around checklist-style written documentation. However, the effectiveness of such an approach to handover has been only anecdotally and analytically demonstrated in the process industries. An empirical, experimental validation of the approach is missing from the literature. This is particularly the case for the industrial process domain where shift handover research has been based primarily on interviews and observation. The present research was motivated by the need to quantify the extent to which shift handover effectiveness can be increased by structuring the verbal shift handover communication around a structured checklist-style written logbook.

2 Experiment Design and Method

A structured shift handover experiment was designed to test the following hypothesis. Using an integrated checklist and logbook to structure the shift handover instead of a less structured logbook will result in:

- A higher percentage of key events and process unit information being communicated during a shift handover.
- An increase in the situation understanding by the second shift operator as he or she takes over control of the unit.

2.1 Within Groups Design

Figure 1 below illustrates the Within Groups experimental design used to compare shift handover performance under each of the two different logbook conditions. Control room operators were assigned to work in pairs. There were two trials, A and B. Trial A always presented the less structured standard logbook condition and Trial B the highly structured checklist-integrated logbook condition. This arrangement of conditions was intended to eliminate any learning effects that would naturally arise if conditions were randomly assigned to either Trial A or Trial B. In other words, we were concerned that if the checklist-integrated logbook were used in Trial A, then shift handover performance with the conventional less structured logbook might appear to be better than it really was. Also, we expected negligible practice effects because shift handover is such a common activity for the experienced operators who were our subjects. So we opted for fixing the order of the two conditions. One of the operators from each pair was randomly assigned to begin Trial A. He/she completed the first half of the incident scenario, resulting in a unit shutdown and then completed a shift handover to the second operator in the pair who took over, executed a unit start-up and completed the scenario. The two operators then switched roles. The second operator completed the first half of a second, different incident scenario, also resulting in a unit shutdown. He/she then completed a shift handover to the first operator who executed a unit startup and completed the second scenario. Thus, each pair of operators completed two shift handovers and in the analysis, the pairs were treated as single entities which completed both structured and unstructured experimental conditions, rather than two individuals, creating the Within Groups design. There were also two scenarios which, for each pair of operators, were counter-balanced across Trials A and Trial B.

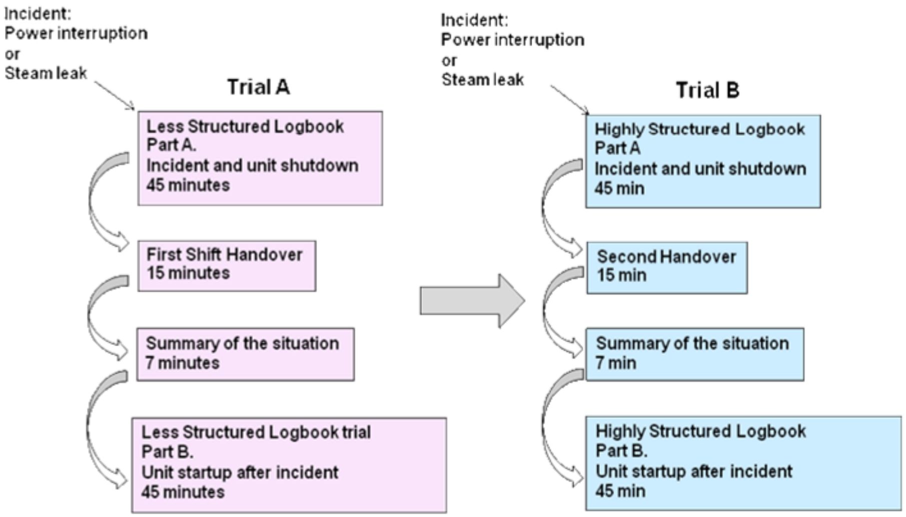


Fig. 1. Block diagram of experimental protocol

2.2 Experiment Method

The experiment was conducted at the ENGEN Refinery (a division of Engen Petroleum Ltd.) in Durban, South Africa. Two versions of the shift handover logbook were evaluated. One was the current less structured logbook in use at Engen's Durban refinery, which featured general headings like "Safety", "Environment" and "Equipment" We deem this as being "less-structured" as the logbook did not specify the details that should be documented, and thus it was up to the outgoing shift operator to decide what should be logged.

The other was an experimental logbook designed around a shift handover checklist developed by an ASM member company based on recommendations from the research on requirements for effective electronic shift logs [3]. Separate from the logbook, the checklist provided details which operators were required to convey during handovers. If a particular section was irrelevant, the operator was required to acknowledge that there was nothing to report under that topic. The experimental logbook design integrated this checklist, resulting in specific sub-headings which the operators had to consider before recording details. We refer to it as the "checklist-integrated" logbook because its headings /categories of information correspond exactly to those used in the ASM member company checklist.

For the experiment, both versions of the logbook were provided electronically to the operators using the Honeywell OMProLog. Both versions of the logbook were considered to be similar in length, as the contents that are to be documented in both versions were identical. That is to say, there were no "missing headings" in the less-structured logbook compared to the checklist-integrated logbook which would otherwise have resulted in intentional omission of detail. This was verified by comparing model logbook entries written by two senior operations engineers who helped in the experiment development. As part of ENGEN's standard operating procedures, the completed logs were printed out to be filed, and the operators were encouraged to use the printed logs during their handovers.

A Honeywell high-fidelity process simulator running the Advanced Distillation Unit Operations Standard Model provided the experimental platform. This model provides a comprehensive and dynamic simulation of typical distillation columns used in gas recovery plants common to most refinery and petrochemical sites. Two failure scenarios were scripted and presented via the simulator:

- Power Interruption- A power interruption occurs, causing pump and fan outage and low tower level alarm, and forcing a unit shutdown
- Steam Line Rupture- A steam line rupture occurs, causing loss of heating steam and forcing a unit shutdown

The events of each scenario were designed to include at least one instance of each information category in the checklist-integrated logbook. The scenario events thus generated a significant number of key items of information that had to be communicated during the shift handover and that affected unit startup during the second shift. Some additional events, not related specifically to the checklist, were included in the scenarios to serve as distractions. Also, the scenarios were designed to force a significant amount of interaction between the console operator, field operators, supervisor, and other plant workers. The field operators, supervisor, and other plant workers were role-played by senior operations engineers from the Engen refinery. They communicated with the console operators during the experimental sessions via radio and mobile phone.

Study participants were operations personnel from the ENGEN Refinery, a division of Engen Petroleum Ltd., in Durban, South Africa. The median age of participants in the study was 39.5 years (range = 27 to 62 years). The median years of DCS experience was 6.5 years (range = 1 to 25 years) and median years in operations was 20.35 years (range = 6 to 35 years). In order to balance the influence of the experienced operator over the inexperienced one, an attempt was made to pair operators that had equivalent experience.

Ten pairs of operators each participated in one 3-hour evaluation session. Each member of the pair had the opportunity to serve as console operator for the first shift one time and as the console operator for second shift one time. As first-shift operator, each member of the pair was responsible for: 1) conducting a failure response and unit shutdown, and 2) preparing and presenting a shift handover. As second-shift operator, each member of the pair was responsible for: 1) understanding the situation at handover, and 2) using that information to conduct a unit startup at the appropriate time. Logbook quality, and second-shift operator recall of situational information were taken as measures of shift handover effectiveness. Operator recall was measured immediately following handover and at various times during the second shift.

3 Results

3.1 Summary of Findings

The results of the experiment supported the hypothesis that using an integrated checklist and logbook to structure the shift handover, instead of a less structured logbook, will increase shift handover effectiveness. Specifically, we found that using the checklist logbook resulted in the following:

- A higher percentage of key events and situational information were documented by the first-shift operator who used the checklist logbook compared to a less structured logbook.
- The second-shift operators who experienced the structured shift handover using a checklist logbook showed an increased understanding of the situation they inherited from the first shift compared to those who experienced a less structured handover.
- The second-shift operators recalled a higher percentage of key items of situational information immediately after shift handover and responded correctly to a higher percentage of probe questions during their shift.
- The above benefits appeared to come at the cost of slightly longer handover times.

3.2 Analysis

The experiment generated a raw score for each team on each of four performance measures (logbook quality, recall, probe responses, and time) for each of the two scenarios. For analysis purposes the raw scores for logbook quality, recall, and probe responses were normalized to percent correct out of the total possible responses. The statistical results reported here used these normalized scores as a way to deal with the fact that the scenarios, which had slightly different numbers of reportable events, were counter-balanced across the two shift handover conditions. Also, all the analyses reported below were exploratory and more liberal statistical significance alpha values and one-tailed directional tests were used [10][1].

Quality of the Logbook. Quality of the logbook was assessed post hoc by reviewing log entries and scoring them against a model logbook generated by operations experts at the ENGEN Refinery. The model logbook represented the information items that the expert would expect in a high quality logbook report for each of the two scenarios used in the experiment. Table 1 shows that the percentage of the total expected information items entered in the logbook was significantly greater for operators who used the Checklist- Logbook than those who used a less structured logbook, F(1, 9) = 6.80, p< 0.02 (one-tailed).

Table 1. Performance summary for logbook quality

Shift Handover Approach	Mean Percent of Total Items Expected	Standard Deviation	% Change
Checklist-Integrated Logbook	76.11	10.48	+18.63%
Less Structured Logbook	57.48	17.64	

[1] Significant alpha values were set at $p < .10$ but marginally significant results ($p < .15$) were also noted.

Ability to Recall Situational Details. A successful shift handover must ensure that the second-shift operator leaves the briefing with a mental model of the situation that is complete, accurate and consistent with the key events that occurred during the first shift. The second-shift operator's ability to recall situational details from the first shift following shift handover was taken as a measure of accuracy and completeness of the second-shift operator's mental model. Table 2 shows a trend toward more complete recall following shift handover among second shift operators that used the Checklist- Logbook compared to operators who experienced shift handover using the less structured logbook , F (1,9) = 2.93, p = 0.12 (one-tailed).

Table 2. Summary of second-shift operator ability to recall first-shift situational information

Shift Handover Approach	Mean Percent of Total Items Recalled	Standard Deviation	% Change
Checklist-Integrated Logbook	51.06	16.24	+8.96%
Unstructured Logbook	42.10	10.34	

Ability to Respond Correctly to Probes. As a further measure of the effectiveness of shift handover communication, second-shift operators were asked a series of probe questions as they worked at starting up the distillation unit. The probes generally requested updates or status of events that had their start or roots in the first shift. Table 3 shows a trend among operators who had experienced a shift handover using the highly structured Checklist Logbook to respond correctly to a higher percentage of probe questions than those who had been briefed with a less structured logbook, F(1,9) = 3.06, p = 0.11 (one-tailed).

Table 3. Summary of second-shift operator ability to respond correctly to probe questions about the first-shift operations

Shift Handover Approach	Mean Percent of Total Probes Responded to Correctly	Standard Deviation	% Change
Checklist-Integrated Logbook	56.43	18.06	+7.86%
Less Structured Logbook	48.57	16.02	

Shift Handover Duration. One of the interesting questions is whether or not using a structured approach to shift handover, such as briefing from the Checklist-Logbook, was more time-consuming than using a less structured briefing approach. Table 4 shows that shift handovers using the Checklist- Logbook took slightly, but not significantly longer than using the less structured shift handovers, F (1, 9) =2.74, p =0.13.

Table 4. Summary and comparison of shift handover duration using the Checklist Logbook to handover using a less structured logbook

Shift Handover Approach	Mean Duration of Shift Handover (in seconds)	Standard Deviation	% Change
Checklist-Integrated Logbook	323.90	80.95	+15.84%
Less Structured Logbook	279.60	90.59	

4 Conclusion and Recommendations

This experiment investigated how the effectiveness of shift handover can be influenced by imposing more structure in the form of a logbook organized around a shift handover checklist. Logbook quality, and second-shift operator recall of situational information immediately following handover and also during the second shift were taken as measures of shift handover effectiveness. Shift handover using the structured, checklist-integrated logbook was compared to handovers that used the legacy, less structured ENGEN Refinery logbook. Although the differences between the two handover approaches were not large, the effect on logbook quality was statistically significant and there were trends toward significant different in the other two measures of effectiveness. Based on these findings, we conclude that providing a sound structure for logging and shift handover, based on key categories of situational information, will likely improve the effectiveness of shift handovers.

While we can only speculate about why larger differences between the two approaches were not observed, it is likely a result of the rather experienced pool of operators who participated in the study. These operators had a median of 20.35 years of operations experience. Even the least experienced operator had spent 6 years in operations. Likewise, these operators were rather experienced in operating a Distributed Control System (DCS), with a median of 6.5 years of DCS experience, which ranged from 1 to 25 years. One would expect that, with all this experience, these operators would have a pretty good idea about what events and information are important to communicate to the second shift, even with little structure imposed on them. We would expect them to perform relatively well in both conditions. Therefore, we can speculate that the increase in effectiveness observed here, though relatively small, is likely a conservative estimate of the benefit of structuring the shift handover around a checklist logbook. Were we to repeat the experiment with less experienced operators, we might expect to find larger differences in performance. Future research needs to address that hypothesis. This question is particularly important in view of the trend in North American refineries toward less experienced personnel in the control room due to retirements in the experienced segment of the workforce. Thus the potential of checklist-structured logging and shift handover to compensate for lack of experience needs to be explored.

One might expect that using the checklist logbook would increase the amount of time required to conduct shift handovers. In the current experiment, shift handovers using the checklist logbook took marginally longer on average than the less structured handovers. However, the time was roughly within the "5-10" minutes often cited as the desired amount of time for shift handover. As one of the participants pointed out in the post-experiment debriefing, filling out the checklist logbook is the most time-consuming aspect of using it. However, it is mostly filled out during the shift and so does not place any significant burden of time on the handover briefing itself. The marginally longer times for structured shift handovers also can be viewed in cost-benefit terms. The small additional amount of time required is more than compensated by the increase in quality of the communicated information. Finally it would be valuable to validate this conclusion by recording the duration of shift handovers in actual operations at a refinery such as ENGEN that has introduced the checklist-integrated logbook into its operations.

Improving communication skills may also be a way to improve shift handovers of both experienced and inexperienced operators. A more controlled investigation of necessary training for communication skill was beyond the scope and resources of this project. However the post-hoc analysis of shift handover verbal interactions revealed that while some effective communication practices were commonly practiced by operators during shift handover, some other practices were not. The latter may provide opportunities to improve shift handover communication through training that focuses on these specific communications skills and practices. Future research should more systematically investigate the extent to which these skills can be trained, what that training should consist of, and what effect training has on shift handover effectiveness. That research also needs to address what is required to maintain over time the new communication skills learned during the initial training.

Acknowledgments. This study was funded by the ASM® Consortium, a Honeywell-led research and development consortium. ASM is a registered trademark of Honeywell International, Inc. The authors would like to acknowledge the contribution of operators at the ENGEN refinery who volunteered to participate in this study. Also we would like to acknowledge the support and contributions of Engen refinery operations staff members, Satish Madhoe and Steve Sanders.

References

1. Bullemer, P., Reinhart, B., Soken, N., Ramanthan, P., Corwin, B.: Towards an understanding of abnormal situation management: Core team site visit summary. ASM Technical Report. ASM Consortium, Minneapolis (1994)
2. Walker, B.A., Smith, K.D., Lenhart, J.E.: Optimize control room communications. Chemical Engineering Progress Magazine, 54–59 (October 2001)
3. Laberge, J., Plocher, T., Goknor, S.: Requirements for Effective Shift Logs. ASM Consortium Technical Report. ASM Consortium, Minneapolis (2006)

4. Cullen, W.D.: The Public Inquiry into the Piper Alpha Disaster, vol. 1& 2. Department of Energy, United Kingdom (1990)
5. British Petroleum: Fatal Accident Investigation Report. Isomerization Unit Explosion: Final Report. British Petroleum, United Kingdom (2005)
6. Parke, B., Kanki, B.: Best practices in shift turnovers: Implications for reducing aviation maintenance turnover errors as revealed in ASRS reports. International Journal of Aviation Psychology 18, 72–85 (2008)
7. Lardner, R.: Safe communication at shift handover: Setting and implementing standards. The Keil Center Ltd, Edinburgh (1999)
8. Eiff, G., Lopp, D., Nejely, D., Vice, M.: Improving Safety and Productivity Through a More Effective Maintenance Shift Turnover (2001), `http://hfskyway.faa.gov`
9. US Department of Energy: Guide to Good Practices for Operations Turnover. Report No. DOE-STD-1038-93 (1993)
10. Dickens, C.D.: Commonsense Statistics. Ergonomics and Design, October 18–22 (1998)

Measuring Self-adaptive UAV Operators' Load-Shedding Strategies under High Workload

Axel Schulte and Diana Donath

Universität der Bundeswehr München (UBM), Department of Aerospace Engineering,
Institute of Flight Systems (LRT-13), 85577 Neubiberg, Germany
{axel.schulte,diana.donath}@unibw.de

Abstract. This article focuses on the experimental identification of changes in human behaviour patterns of UAV-operators guiding multiple UAVs from a helicopter cockpit. These changes are based on self-regulation mechanisms of the operators to adapt to the current task and workload demands. Main objective of the use of these so called self-adaptive strategies is to avoid overload situations, and to retard exceeding capacity limits, to maintain overall acceptable performance as long as possible. Expressed by shedding and deferring tasks of lesser importance, or the relaxation of self-imposed criteria, these strategies lead to an observable change of human behaviour patterns, prior to grave performance decrements. This article describes a laboratory experiment utilising a virtual flight simulator to stimulate operator's workload and observe their mitigation strategies by means of gaze detection and a detailed interaction monitoring. Using the observed behaviour changes in an assistant system as indicator for high workload situations of the operator, it shall be possible to support the operator prior the occurrence of errors.

Keywords: multi-UAV guidance, subjective workload, self-adaptive strategies, human behaviour model, eye movements.

1 Introduction

UAV-operators are predominantly faced with supervisory control tasks, comprising the control of UAVs as such, the analysis of sensor images provided by the UAVs and the classification of detected objects, i.e. the tactical situation management. Increasing the number of UAVs guided by a single UAV-operator, results sooner or later in an overload situation for the human, due to the limited available attention resources, which need to be shared between the numerous tasks associated to the guided UAVs. In such situations the human is not able to compensate any further demands by investing more effort, so any further increase of task load often results in a decrease in operator performance and therefore in performance decrease of the overall system. To avoid such overtaxing situations for the human an assistant system needs, among other things, the ability to detect if the human operator is currently overtaxed [1]. To provide support at an early stage, theses systems should recognise such overtaxing situations of the human, prior to the occurrence of errors, thereby facing the challenge, that there are no well defined criteria or measurable values, indicating the limit

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 342–351, 2011.

between high workload and overtaxing. Furthermore the underlying notion of workload can be seen as a multidimensional, psychological, subjective phenomenon which is inaccessible to a direct measurement. Several different approaches try to infer human operator workload by the use of observable parameters. These approaches range from methods solely based on estimations up to methods, which continuously access performance (e.g. reaction time, error rate) or psycho-physiological data (e.g. EEG, pupil diameter) [2]. Depending on the method, their theoretical underpinnings and measures used, these approaches either act proactively or reactively, i.e. using feed forward estimations on the basis of the current task load, but without any consideration of the actual operator response, or providing a deviation or error correction based upon the detection of certain human response patterns [3].

2 Behaviour Based Approach

We propose an approach aiming at the detection of critical operator workload prior to the occurrence of errors, or grave performance decrements. We use certain sets of continuously accessible human performance and behaviour parameters. Therefore, we observe the manual and visual interactions of the operator with the technical system during task accomplishment. This allows to gain insight in how (and not just how well) the human operator accomplishes the given tasks. We hypothesise to detect the self-regulating mechanisms of the human in high workload situations, while trying to adapt themselves to the current task and workload situation. In literature these mechanisms are referred to as self-adaptive strategies [4].

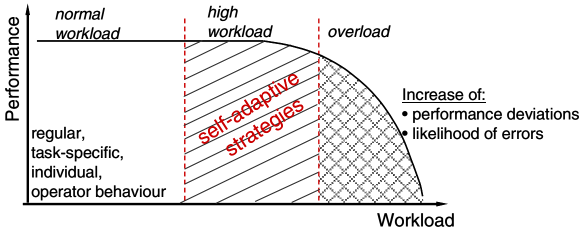


Fig. 1. Self-adaptive strategies occur prior to performance decrements

Applying these strategies, tasks of lesser importance will be shared, shed or deferred, which leads to an observable behaviour change. In general, there can be distinguished two fundamental classes of self-adaptive strategies, i.e.

- *load-sharing strategies*, i.e. the transfer of tasks to other team members or to functions of the automated system, and
- *load-shedding strategies*, i.e. changing the way a task is accomplished.

In the first case, the applicability of load-sharing depends upon the availability of automation or other team members. In the second case, load-shedding depends upon the flexibility (i.e. the degrees of freedom) the task itself provides. In both cases, the selection of the strategy is dependent on the operator's available work capacity.

In the case of *load-shedding strategies* the tasks will be accomplished in a more economic, not necessarily perfect way. Primary objectives will be pursuit at the expense of secondary objectives. This leads to the observable behaviour adaptations such as task prioritisation, disregard of subtasks, change in task accomplishment, or altered attention allocation [5].

Main objectives for the use of these strategies are to avoid overload situations by keeping the workload within bearable limits [6] and to retard possible capacity limits as long as possible to maintain overall performance [7][8]. Since increasing task load will lead to a progressive change in human behaviour of a human operator, the observation of the behaviour can be used as an indicator for workload.

To use these self-regulating mechanisms of humans as trigger for supporting functions in future assistant systems, human behaviour models are required, representing human behaviour within normal and high workload conditions. Hence, we need to consider that human behaviour is

- *individual*, as a consequence of e.g. skills, abilities and training,
- *task specific*, since each task or task combination implies a certain set of interactions (e.g. manual, visual), and
- *dependent on the current perceived subjective workload situation of the human*, referring to the change of human behaviour as a consequence of the use of self-adaptive strategies.

3 Experimental Operator Behaviour Acquisition

To gather and investigate human behaviour of UAV-operators in the accomplishment of their supervisory control tasks within different workload conditions (normal and high workload), extensive simulator trials were performed.

3.1 Experimental Design

The experiments, performed within a fixed based research helicopter and multi-UAV simulator, referred to a MUM-T scenario, i.e. the guidance of multiple UAVs by a human UAV-operator located in a helicopter cockpit. Here, UAVs were used as remote sensor platform for real-time reconnaissance during a simplified military air assault mission (Fig. 2). The UAVs were guided along pre-planned routes (*FMS-based*), which could be adapted to the current situation by the operator at any time. The main tasks of the UAV-operator were the guidance of the UAVs as such, the analysis of sensor images, taken by the UAVs and the classification of recognised objects. Therefore, the UAVs were equipped with a thermal camera and a video data link. The total time of the experiment was 95 minutes. To provoke the occurrence of self-adaptive strategies as a consequence of an increase in task load and subjective perceived workload of the operator, the task load was systematically increased over the course of mission by the introduction of embedded secondary tasks (i.e. mission re-planning, threat localisation), as well as by the increase in the number of UAVs to be controlled by the UAV-operator within two consecutive missions (Fig. 2, right side). To get familiarised with the UAV-operator station layout and the handling of the FMS-based guidance of UAVs, as well as with the reconnaissance task, the subjects got one full day of training.

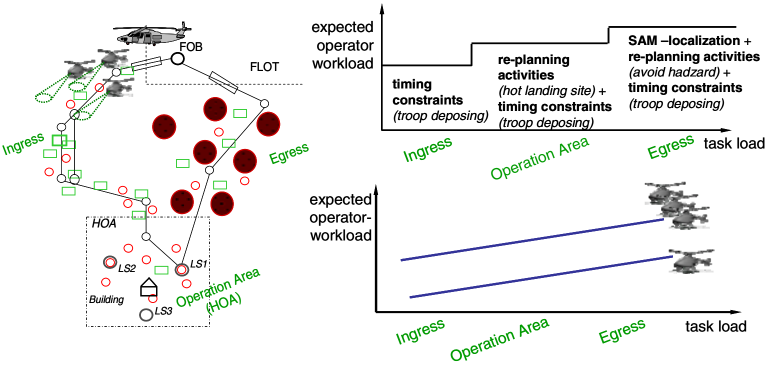

FOB
FLOT
Ingress
Egress
HOA
LS2
LS1
LS3
Building
Operation Area (HOA)
expected operator workload
timing constraints (troop deposing)
re-planning activities (hot landing site) + timing constraints (troop deposing)
SAM –localization + re-planning activities (avoid hadzard) + timing constraints (troop deposing)
Ingress
Operation Area
Egress
task load
expected operator-workload
Ingress
Operation Area
Egress
task load

Fig. 2. Scenario (left), increase of task load during mission (right, top), increase of task load over two consecutive missions (right, bottom)

To capture human behaviour during task execution the simulator was equipped with faceLAB, a contract free, video-based eye movement measurement system. Furthermore, manual interactions of the UAV-operator, e.g. button presses were recorded. To get a relationship between the observed operator behaviour and his subjective perceived workload subjective workload ratings (NASA-Task Load Index) were performed in discrete intervals. In addition the operator performance was captured by the use of the following performance parameters: the number of classified objects, the required time for the accomplishment of the object-identification task, classification errors, operating errors. The subjects were four military helicopter pilots, two of them average experienced (around 550 flight hours, 150 hours as commander, around 30 yrs.) and the other two highly experienced (around 1700 flight hours, 1550 hours as commander, around 42 yrs.); the latter ones were well-trained flight instructors.

3.2 Investigated Task

The overarching tasks of the UAV-operator were the reconnaissance of the ingress and egress route and the Helicopter Operation Area (HOA) including the landing sites and a building located within the HOA. The focus for the behaviour investigation was the object identification task, which is a self-contained, repetitive subtask of the superior route-reconnaissance task. To get sufficient behaviour data, a great number of objects (friendly and foe ones) to be recognised and identified by the operator, were placed along the course of the mission (Fig. 2, left side). The task can be further divided into three subtasks:

- *"recognise and tag"* is the phase between the recognition of a hotspot, and the tagging of the hotspot in the map,
- *"classify"* starts with centring a live video-streaming sensor of the UAV on the hotspot, followed by the classification (civil or hostile) of the object,
- *"insert result"* is the insertion of the result into the interactive map.

Each of these subtasks can be broken down into several sub-subtasks, which are characterised by certain observable manual and visual interactions of the operator (for further details see [9]). Fig. 3 shows the UAV-operator control station, consisting of two touch-displays and the available display modes.

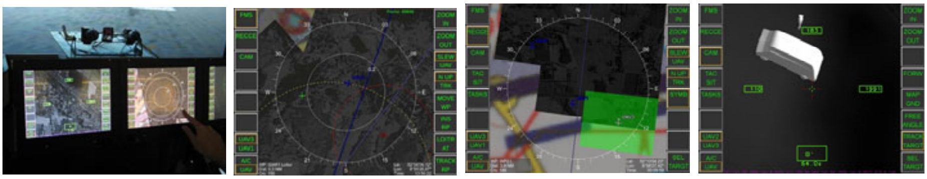


Fig. 3. Operator control station (left), available display modes

4 Experimental Findings

The following sections discuss the findings concerning performance parameters, subjective workload, and self-adaptive strategies of the human operator.

4.1 Operator Performance and Subjective Workload

In brief the following observations concerning mission performance were made during the experiments. The *relative number of detected objects* slightly went down from the Ingress to the HOA mission phase and fully dropped on Egress. The latter effect will be further discussed in the following sections. Managing three UAVs as opposed to one decreased the number of detected objects as well. *Errors in object classification* could almost not be observed throughout the missions and subjects. *The time required for the full object identification task* increased only slightly over the course of mission as well as in the guidance of three UAVs. The *number of errors in the handling of the system* varied along the independent parameters mission phase and number of UAVs. To sum up, the operators overall maintained their performance on a good level, although first slight performance decrements could be observed.

Subjective workload ratings (NASA TLX) were collected in discrete intervals. Independent of the guidance of one, or three UAVs, or over the course of mission there was only a very slight increase of subjective workload of all operators. In summary, we could observe that despite of the massive increase of task load, no explicit increase in the subjective perceived workload could be registered. The results of the subjective ratings hardly allow an inference of overtaxing situations.

4.2 Self-adaptive Strategies

As shown before the expected increase in the perceived subjective workload of the subjects due to the massive increase in task load could not be observed. Only slight performance decrements in the task accomplishment could be noticed. The overall success in the classification of attended objects remained quite stable. Instead, the human operators responded with a various behavioural changes, which were not introduced consciously at all times. By use of those load-shedding strategies, the operators tried to keep their subjective workload within bearable limits. The observed adaptations of the subjects' behaviours within the studied task were:

1) *proactive task reduction*, i.e. the sole use of only one of the available UAVs for the route reconnaissance task,
2) *less exact task performance*, i.e. consciously accepting an ambiguous live video stream of the object, insufficient to make a secure classification,
3) *omission of subtasks*, i.e. cutting down certain operating steps,
4) *complete neglect of object identification task* during entire mission phases,
5) *purposeful delay of task accomplishment,* i.e. the intended interruption of tasks in highly demanding situations, and the continuation of this task in lower workload situations.

It appears that some of these strategies occurred only once, however other strategies could be observed more frequently, with more than one subject, or in different mission segments.

Table 1. Load-shedding over mission phases and number of UAVs

load-shedding stategies.	***subject 1***	***subject 2***	***subject 3***	***subject 4***
task reduction	3 UAV, Ingress			
less exact perf.				1 UAV, Ingress 3 UAV, HOA
omission of subtasks	1 UAV, HOA 3 UAV, HOA	1 UAV, Ingress 1 UAV, HOA		1 UAV, Ingress 3 UAV, HOA
neglect of task	1 UAV, Egress 3 UAV, Egress	1 UAV, Egress 3 UAV, Egress	1 UAV, Egress 3 UAV, Egress	1 UAV, Egress 3 UAV, Egress
purposeful delay		1 UAV, HOA		

Table 1 shows the above mentioned load-shedding strategies, for each individual subject. The table entries show the number of UAVs and the mission segment.

Proactive task reduction. The overall task of the UAV-operator was the reconnaissance of the helicopter route (ingress and egress, as well as the operational area) by the use of all available UAVs, either one or three. Therefore, the operator had to alternately point his/her attention to the sensor images of the different UAVs in order to search for target candidates (i.e. the hotspots). This was done as expected by the operator within the initial mission segment, ranging from the FOB (start point of HC and UAVs) up to the FLOT, to be crossed via an ingress corridor. With entering the corridor by UAV1, one operator completely dropped allocating attention to the remaining UAVs (UAV2, UAV3) as well as to their sensory output. From this moment the operator only focused his attention to UAV1 (see Fig. 4, left). Fig. 4 (right) shows the attention allocation (measured by fixation dwell times) of the operator to the sensor images of the different UAVs prior to and after entering the ingress corridor delineated by the dashed line. By the use of this strategy the operator proactively reduced his tasks, trying to avoid reaching his capacity limits in terms of attention resources and therefore avoided the possible upcoming of an overtaxing situation.

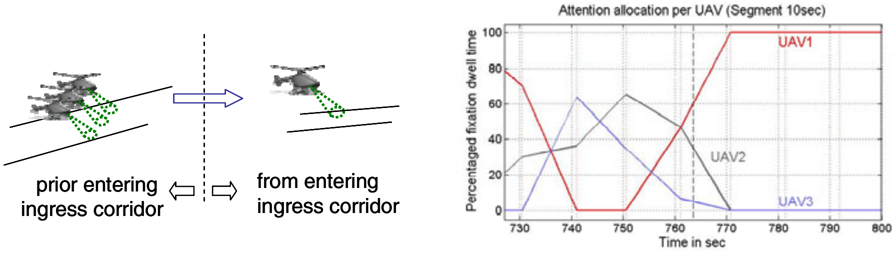


Fig. 4. Load-shedding by the proactive task reduction visible in the attention allocation of subject 1 to the different UAVs

Less exact task performance. This strategy refers to an observed behaviour change of operators solely within the ”classify” phase of the studied task. For the accomplishment of this subtask the subjects had to view the live video stream of the currently selected UAV, which was previously aligned to the hotspot to be classified. To clearly classify the object, the operator needs to select the maximum available zoom. Furthermore, there needs to be assured that the distance between the UAV and the object is less than 1 nautical mile.

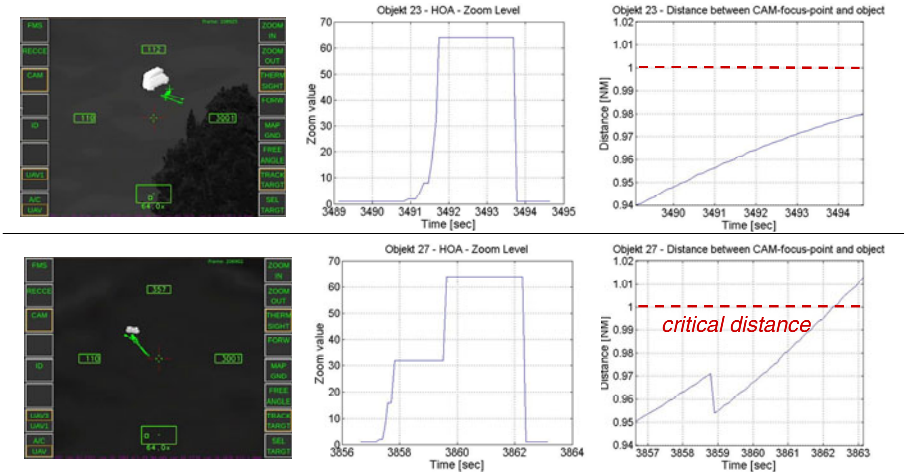
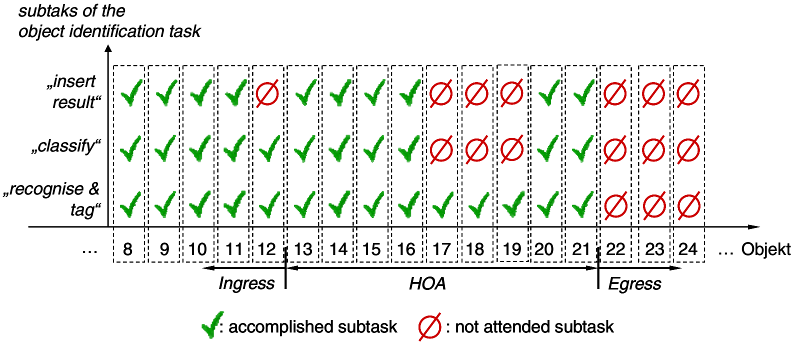


Fig. 5. Classification subtask under normal conditions (top), modified as a consequence of load-shedding (bottom)

Beyond that distance there is no unambiguous classification possible anymore. In this case the operator has to reduce the distance between the UAV and the object again, which is oftentimes accompanied with a laborious re-planning of current UAV-flight-path. Fig. 5 shows two different examples of the accomplishment of the *classify* subtask, depicted under normal conditions as well as under the use of a load-shedding strategy. Both show the received sensor image, the zoom-factor for display of the object and the distance between the UAV and the object. The load-shedding strategy in this case implies the operator accepting the insufficient image as a basis for the object classification. As a consequence, the operators were only able to guess the classification of the objects, resulting in an increased threat for the mission in case of a wrong classification of a hostile object.

Omission of subtasks. Usually the object-identification task consists of the three subtasks *"recognise and tag"*, *"classify"* and *"insert result"*. In demanding and often time critical situations, such as the reconnaissance of possible landing sites within the HOA, we observed that operators omitted some of these subtasks to a certain extent.



	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
„insert result“	✓	✓	✓	✓	∅	✓	✓	✓	✓	∅	∅	∅	✓	✓	∅	∅	∅
„classify“	✓	✓	✓	✓	✓	✓	✓	✓	✓	∅	∅	∅	✓	✓	∅	∅	∅
„recognise & tag“	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	∅	∅	∅
...	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24 ... Objekt

Fig. 6. Omission of subtasks of subject 1 managing three UAVs

This strategy is depicted in Fig. 6 for subject 1 managing three UAVs. Generally the omission of the subtask *"insert result"* is a consequence of a previously started but failed classification of the object, which the operator did not want to redo at a later time. The complete drop of the classification process (consisting of the subtasks *"classify"* and *"insert result"*) however is a phenomenon which occurred repeatedly in time critical situations. According to statements of operators during the debriefings, the omission of these subtasks was due to an attempt to avoid a critical overload situation. In the cases 17-19 (in Fig. 6) the operator informed his team-member (the pilot of the helicopter) to pay attention while passing the objects on his way to the landing site. Here the operator not simply omits the accomplishment of the subtasks, but he shifts the responsibility for a safe flight of the helicopter to the pilot. This behaviour might be classified as a *load-sharing* strategy.

Complete neglect of the object identification task. A complete drop of the object identification task including the search for hotspots could be observed for all subjects and in all configurations on the egress part of the mission (Fig. 7, left). During the egress phase the task load had been further increased by introducing SAM sites at unknown position. Detection and localisation of those threats created an additional embedded, secondary task. Due to their imminence it was expected that these secondary tasks for the time from a radar contact to the determination and entry of the position would become the primary task. However, it appears, that also in times, which were free of any hazard for the UAVs and the helicopter the subjects almost never resumed to the search for hotspots. Fig. 7 (right) shows the percentage of use of the different map range circles (i.e. map scale) over the course of mission. Only range circle values of 0.2 and below allow the detection of hotspots whereas the detection with range circle values above 0.2 is nearly impossible. As depicted in Fig. 7, during the egress-part of the mission the UAV-operator (subject 4) used the required range circle value (≤ 0.2) only in a few percent of the time (here referred to the time free of any hazards). This indicates that the UAV-operator did almost never resume the search for possible targets.

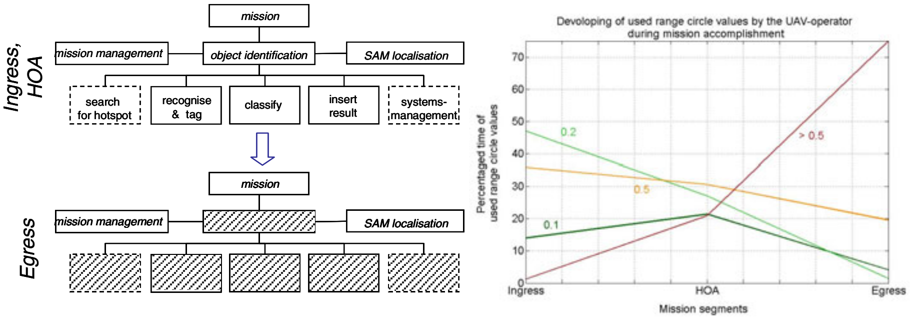


Fig. 7. Complete drop of the object-identification task of all subjects in both configurations (1 or 3 UAVs) during egress

Purposeful delay of task accomplishment. This strategy refers to an observed situation during the reconnaissance of the landing three sites in the HOA using one UAV. This situation is principally characterised as a time critical one, since the UAV-operator has to inspect all possible landing sites as fast as possible and thereupon direct the helicopter to a secure landing site.

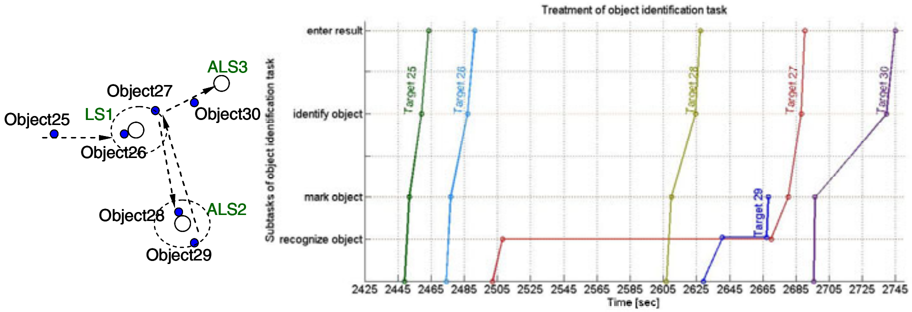


Fig. 8. Load-shedding through delay of task accomplishment and prioritising of tasks during time critical situations while landing site inspection

As shown in Fig. 8 the operator initially detects two hostile objects (Object 25, Object 26) on the way to landing site LS1, which implies that LS1 is "hot" and therefore inappropriate for landing. Then the operator detects Object 27, which could be proven by several, consecutive fixations on the hotspot followed by several fixations on the button, to initiate the tagging. As this object (Object 27) is close to the landing site LS1, which at that time already had been classified as hot, *the operator deferred this task*, and prioritised the reconnaissance of the route to landing site ALS2 first. Thereby, he encountered Object 28, which was immediately classified by the operator as foe, and therefore ALS2 as "hot" too. At the same time, the operator detects Object 29. Since ALS2 was also already classified as "hot", the classification of that object was irrelevant for the operator. Therefore, the operator immediately omits the classification and the insertion of the result for this object. Instead of that the operator focuses on the more urgent task, which means the reconnaissance of the route to landing site ALS3. Thereby, the operator passes the previously deferred Object 27, which he now processes. In this case it appears that the operator systematically defers the accomplishment of tasks in time critical situations, and prioritises the processing of other, in this situation more relevant tasks.

5 Conclusions and Perspectives

Experiments have shown that operators kept their workload within manageable ranges, while their performance only slightly decreased. This was achieved by the application of self-adaptive strategies of operators in high demanding work situations. During the experiments, several different load-shedding strategies could be observed. Next steps will be the development of computational models on the basis of the identified, quantifiable parameters. Using these models within an assistant system, an operator-adaptive support will be possible.

References

1. Onken, R., Schulte, A.: System ergonomic Design of Cognitive Automation – Dual Mode Cognitive Design of Vehicle Guidance and Control Work Systems. Springer, Heidelberg (2009)
2. Parasuraman, R., Hancock, P.A.: Mitigating the Adverse Effects of Workload, Stress and Fatigue with Adaptive Automation. In: Hancock P.A., Szalma, J.L. (ed.) Performance under Stress. Ashgate (2008)
3. Scerbo, M.W.: Theoretical Perspectives on Adaptive Automation. In: Parasuraman, R., Mouloua, M.(ed.): Automation and Human Performance-Theory and Applications. LEA (1996)
4. Canham, L.S.: Handbook of Human Factors Testing and Evaluation. Operability Testing of Command, Control & Communications in Computers and Intelligence (C41) Systems. Mallory International (2001)
5. Veltman, J.A., Jansen, C.: The role of operator state assessment in adaptive automation. TNO-DV3 2005 A245 (2006)
6. Sperandio, A.: Variation of operator's strategies and regulating effects on workload. Ergonomics 14, 571–577 (1971)
7. Sperandio, A.: The regulation of working methods as a function of workload among air traffic controllers. Ergonomics 21(3), 195–202 (1978)
8. Parasuraman, R., Rovira, E.: Workload Modelling and Workload Management: Recent Theoretical Developments. Army Research Lab ARL-CR 0562 (2005)
9. Donath, D., Rauschert, A., Schulte, A.: Cognitive assistant system concept for multi-UAV guidance using human operator behaviour models. In: HUMOUS 2010 (2010)

Display Requirements for an Interactive Rail Scheduling Display

Jacqueline M. Tappan[1], David J. Pitman[1], Mary L. Cummings[1], and Denis Miglianico[2]

[1] Humans and Automation Laboratory, Massachusetts Institute of Technology,
77 Massachusetts Avenue, 37-311, Cambridge Massachusetts, 02139 USA
[2] Alstom Transport, 48 Rue Albert Dhalenne, 93400 Saint Ouen, France
jtappan@mit.edu, edave@mit.edu, missyc@mit.edu,
denis.miglianico@transport.alstom.com

Abstract. This work, a collaboration between Alstom Transport and the MIT Humans and Automation Laboratory (HAL), is focused on the development of an interactive in-cab scheduling interface for train operators. Currently, operators rely on a combination of paper schedules, paper speed charts, and rote memorization to meet the many demands of train operation. The separation of this information over multiple sources shifts driver attention away from the windscreen and may result in increased workload levels and safety compromises. A Hybrid Cognitive Task Analysis (hCTA), which derives the information requirements necessary to meet mission goals directly from operational tasks, was conducted to generate cognitive requirements for the desired scheduling display. The resulting seventeen requirements were used to guide the development of a new scheduling display, which is presented.

Keywords: Information requirements, cognitive task analysis, rail, schedule management, decision ladders.

1 Introduction

The train operation environment has become increasingly complex. To operate a train from station of departure to station of arrival, an operator must gather information from multiple digital displays and paper supplements, while monitoring the external environment through the windscreen. A particularly difficult task for train operators is scheduling. Currently, operators rely on a combination of paper schedules, paper speed charts, and rote memorization to perform the scheduling task. The separation of this information (e.g., arrival times, speed restrictions, voltage changes) over multiple sources results in the shifting of driver attention away from the windscreen and may increase workload levels and compromise safety. The combination of the required information on a single display would ease workload levels and potentially result in safer train operations. In addition, designing a display that takes advantage of information visualization using both textual and graphical elements would enhance

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 352–361, 2011.

driver adherence to posted schedules. This design could also include information on speed profiles to optimize trip efficiency [1].

This work, a collaboration between Alstom Transport and the MIT Humans and Automation Laboratory (HAL), is focused on the development of a minimum information interface for increased efficiency and safety during rail operations through the development of an interactive in-cab scheduling interface. Information about current high speed rail operations in France was collected through three train trips: two trips on RER (Réseau Express Régional) train routes and a single trip on a TGV (Train à Grande Vitesse) train traveling from gare de L'Est to Strasbourg. A visit to the Alstom Transport simulator located in Belfort also provided observation opportunities. These observations revealed that schedule management was indeed an existing problem for train operators.

The Hybrid Cognitive Task Analysis (hCTA) [2] method was used to generate the cognitive requirements for the entire train operation process. The requirements specific to train scheduling were then used to guide the development of a scheduling display. The next section will detail the hCTA process. The final prototype resulting from the generated requirements will then be presented.

2 Hybrid Cognitive Task Analysis

The goal of the hCTA process is to generate the functional and information requirements for an interface design within a complex system, starting from a high-level scenario task description [2]. The hCTA process attempts to define the workflow of a human operator within a complex environment, deriving a complete set of computer interface information requirements necessary to meet mission goals directly from operational tasks. The hCTA consists of the following components: 1) Scenario Task Overview, 2) Event Flow Diagrams, 3) Situation Awareness Requirements, 4) Decision Ladders (and jointly, display requirements), and finally, 5) Information and Functional Requirements. hCTAs have been utilized in a variety of domains, including in the design of displays for the control of unmanned underwater vehicles (UUVs) [3] and submarine surface collision avoidance [4].

An overview of the hCTA process for this interactive rail scheduling display effort is discussed in the following subsections, as well as the resulting output from each step for the train operation process. The full hCTA developed for this research can be found in Tappan, Pitman, et al. [5].

2.1 Scenario Task Overview

A scenario task overview formalizes the mission statement for a complex work environment into a set of distinct phases and tasks, similar to a hierarchical task analysis [6]. A phase represents an abstract grouping of similar tasks designed to meet some common goal, and for the purposes of this effort, phases can be temporally defined. Implicitly, each phase has a set of sub-goals that the operator is trying to achieve while engaged in that phase of the scenario. Often, these sub-goals are represented as specific tasks for a phase. Phase tasks are not limited to just physical actions; they may also include temporal vigilance tasks (i.e., monitoring gauges) or problem solving (i.e., determining new speed to stay on schedule).

Ultimately, the scenario task overview provides a basis for the rest of the hCTA analysis by transforming a qualitative description of an operator's job into a set of quantifiable tasks, which are then precisely defined using event flow diagrams.

For this project, four phases were identified relating to a train operator's mission of driving a train along a route, with intermittent stops at stations: *Before Departure* (BD), *Leaving Station* (LS), *En Route* (ER), and *Arrival at Station* (AS). In the first phase (BD), the operator prepares the train cab for departure. In the second phase (LS), the operator guides the train out of the station and accelerates to the first posted speed limit. In the third phase (ER), the driver continuously monitors the displays and gauges within the train cab to ensure that all systems are within acceptable bounds and are functioning correctly. In the final phase (AS), the driver decelerates into the arrival location and brings the train to a complete stop. In all, 47 subtasks were identified within the four phases.

2.2 Event Flow Diagram

An event flow diagram represents the temporal constraints (i.e., when and in what order) of events and tasks that occur within a specific phase of a mission. There are three basic types of symbols in an event flow diagram: 1) Processes, which require interaction between the operator and system (e.g., activate control *x*), 2) Decisions, which require the operator to make rule-based (simple) or knowledge-based (complex) judgments, and 3) Loops, which are processes that occur iteratively until a pre-determined event occurs (e.g., monitor surroundings for *x*). Three additional symbols often used in event flow diagrams are Phases (representing other flow diagrams accessible from the current diagram), Assumptions (information or requirements assumed to have been met before phase execution), and the transition arrows linking all diagram elements. The event flow symbols are shown in Figure 1.

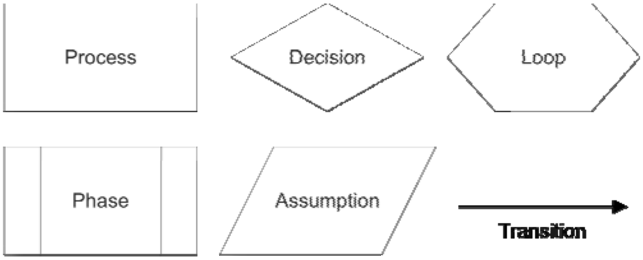


Fig. 1. Event flow diagram legend

It is important to note that within the flow diagrams (and independent of which loop the operator is in), it is always possible for the operator to follow a preemption path where the operator halts their current task and instead attends to another event that has become more urgent.

A single event flow diagram was created for each of the four phases. Alphanumeric labels were given to the blocks so that they could be cross-referenced throughout the rest of the hCTA process (P for processes, D for decisions, L for loops). In all, 92 processes, 13 decisions, and 31 loops were identified over the four operational phases.

The work of train operators is predominantly comprised of the continuous monitoring of their environment, both within the train cab and through the windscreen, with multiple monitoring tasks occurring simultaneously. This monitoring task can be depicted using basic event flow symbols by combining a *loop* symbol, representing the item being monitored, and a *process* symbol, representing the method of monitoring. Upon detection of a signal, the *loop* would be exited, leading to a set of *decisions* and/or *processes* before returning to the original *loop*. While the monitoring task can be represented using these basic symbols, it is difficult to depict multiple concurrent monitoring tasks due to the temporal connections that dictate movement through flow diagrams. In order to overcome this limitation, a new graphical symbol was created for the train event flow diagrams, termed the *monitoring block* and represented visually by a dotted outline. This outline is placed around each monitoring task, grouping the symbols within, and is labeled with the monitoring task that is represented by the symbols. These blocks are not directly connected to the rest of the event flow diagram, conveying their continuous nature throughout the duration of each phase. In the train environment, monitoring blocks occurred within all phases, and included monitoring the radio, weather, and speed.

A portion of an event flow diagram, depicting the continuous monitoring by the train operator for threats on or beside the track, is shown in Figure 2. The dotted outline surrounding the flow diagram symbols indicates that this is a *monitoring block*, with the title *threat detection* conveying its purpose. The block begins with the loop, *L15: Monitor surroundings for pedestrians, tunnel.* If the exit condition for the loop is not met, the operator continues to gather environmental cues through the windscreen (P40). If an object is detected, the loop is exited and the operator must decide whether the object is on or beside the track (D4). If the operator decides that the object is on the track, tasks P41 and P42 must be completed. If instead the object is deemed to be beside the track, task P43 must be completed. The operator then resumes the original loop task until a new threat is detected.

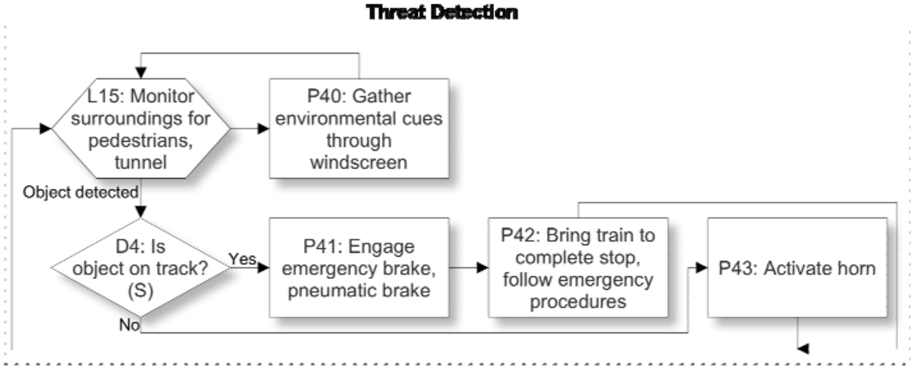

Threat Detection
L15: Monitor surroundings for pedestrians, tunnel
P40: Gather environmental cues through windscreen
Object detected
D4: Is object on track? (S)
Yes
No
P41: Engage emergency brake, pneumatic brake
P42: Bring train to complete stop, follow emergency procedures
P43: Activate horn

Fig. 2. Portion of an event flow diagram

2.3 Decision Ladders

Decision ladders are tools that aid in capturing the states of knowledge and information-processing activities necessary to reach a decision [7]. In the hCTA process, decision ladders are created for each complex decision identified in the event flow diagram process to better understand the information required to adequately support the human decision-maker when faced with such a decision. Complex decisions are those that include many dynamic variables and occur in uncertain environments. These are in contrast to simple decisions (also identified in the event flow diagram), which are typical binary decisions (i.e., yes vs. no), and can be easily made from information readily available in the environment.

Decision ladders depict the decision-making process, beginning with the observation of an anomalous state, the identification of that state, the interpretation and evaluation of the ultimate goal in addressing the decision, and finally, the determination and execution of the correct response. In a decision ladder, this process is categorized using three levels of human behavior: 1) Skill-Based Behavior, or unconscious control, 2) Rule-Based Behavior, where decision-making is based on stored rules learned from previous experience, and 3) Knowledge-Based Behavior, where decision-making is based on environmental cues and individual goals [8].

Once a primary decision ladder is completed, two iterations are produced: 1) a ladder incorporating display requirements, and 2) a ladder incorporating potential levels of automation [9]. The display requirements and automation levels are listed in annotations beside the related information-processing activity. These annotations detail data that needs to be displayed in order for the human decision-maker to progress to the next stage, or the role that automation should play in the human's information processing.

The resulting three decision ladders, along with the generated situation awareness requirements (SARs), guide the next step of the hCTA, the development of information and functional requirements. These requirements are then used to begin the interface design process.

The single decision ladder developed for the rail domain was for the complex decision *is train ahead or behind schedule?*, which originated from the continuously monitored *Speed* loop in the *En Route* flow diagram. This decision ladder depicts the train operator continuously monitoring the current and estimated arrival times. When a schedule anomaly is detected, the operator determines the extent to which the train is behind or ahead of schedule, concludes whether some action will resolve the anomaly, and if so, takes the required action to return the train to on-time travel. The display requirements (Figure 3) and potential automation levels to assist the operator through this decision-making process were included in two iterations of the original decision ladder.

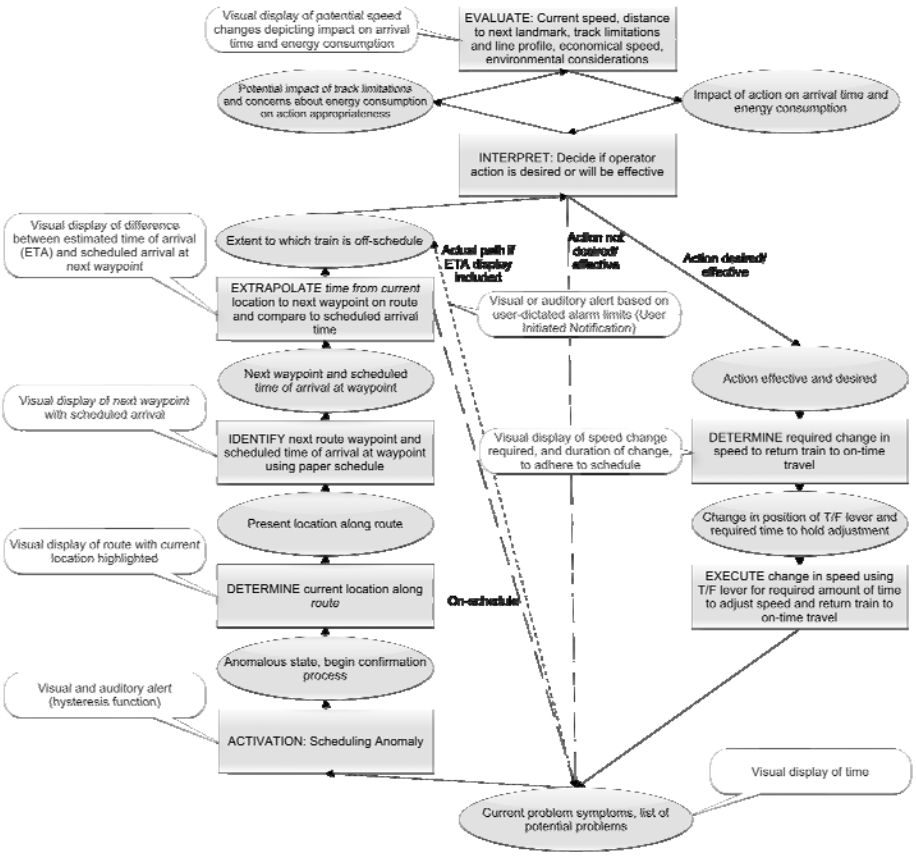

Visual display of potential speed changes depicting impact on arrival time and energy consumption
EVALUATE: Current speed, distance to next landmark, track limitations and line profile, economical speed, environmental considerations
Potential impact of track limitations and concerns about energy consumption on action appropriateness
Impact of action on arrival time and energy consumption
INTERPRET: Decide if operator action is desired or will be effective
Visual display of difference between estimated time of arrival (ETA) and scheduled arrival at next waypoint
Extent to which train is off-schedule
Actual path if ETA display included
Action not desired/ effective
Action desired/ effective
EXTRAPOLATE time from current location to next waypoint on route and compare to scheduled arrival time
Visual or auditory alert based on user-dictated alarm limits (User Initiated Notification)
Next waypoint and scheduled time of arrival at waypoint
Action effective and desired
Visual display of next waypoint with scheduled arrival
IDENTIFY next route waypoint and scheduled time of arrival at waypoint using paper schedule
Visual display of speed change required, and duration of change, to adhere to schedule
DETERMINE required change in speed to return train to on-time travel
Present location along route
Change in position of T/F lever and required time to hold adjustment
Visual display of route with current location highlighted
DETERMINE current location along route
On-schedule
EXECUTE change in speed using T/F lever for required amount of time to adjust speed and return train to on-time travel
Anomalous state, begin confirmation process
Visual and auditory alert (hysteresis function)
ACTIVATION: Scheduling Anomaly
Visual display of time
Current problem symptoms, list of potential problems

Fig. 3. Decision ladder with display requirements

2.4 Situation Awareness Requirements

After completing the event flow diagrams and in conjunction with developing the decision ladders, Situation Awareness Requirements (SARs) are generated. Decision ladders represent a specific known decision process. However, for the majority of supervisory control tasks, operators are essentially monitoring the system to detect some anomaly or need for intervention, which may not be clearly mapped to a decision. SARs are generated for these tasks.

Situation awareness is commonly split into three levels, Perception (Level 1), Comprehension (Level 2), and Projection (Level 3) [10]. During Level 1, the human operator perceives any available information from the environment, (e.g., visual, auditory, or tactile data). During Level 2, the human operator integrates the acquired data to form and guide his or her current mental model of the environment. Finally, during Level 3, the human operator forecasts future situation events based on his or her current mental model, allowing for timely and accurate decision making.

SARs were generated for each of the four mission phases. Each requirement is directly linked to at least one process, loop, or decision from the event flow diagrams with this link included beside the requirement in the table. For example, for the operator to determine and select a voltage level (P10 in the *Before Departure* event flow diagram), he or she would need a list of available voltages and the required voltage level. Information about upcoming voltage changes would also be useful in order to prepare for future changes. Therefore, a resulting Level 1 (Perception) SA requirement in the *Before Departure* phase was *available voltage levels*. The related Level 2 (Comprehension) SA requirement was *required voltage level for current location*. Finally, the associated Level 3 (Projection) SA requirement was *voltage changes during route traversal*. A total of 45 situation awareness requirements were identified.

2.5 Information Requirements

The resulting SARs, in combination with the already-produced display requirements from the decision ladder, are used to populate the final list of information requirements. These requirements are sorted into functional groupings based on the functions they support. Within the Information Requirements (IR) table, the source of the requirement is listed, allowing for the requirement to be traced to previous portions of the hCTA and therefore justifying the need for the requirement. The ability to trace requirements is critical because if one requirement is not included in the final display design (typically for cost or implementation concerns), the impact of such a decision can be assessed across the entire system.

As a result of the hCTA, 58 information requirements were defined and then grouped into three display functional groupings. The first display grouping, Situation Awareness Display (SAD), would transmit general status updates about overall system and sub-system operation. The IRs important for the detection and resolution of system errors were grouped into the second category, Error Identification and Recovery Display (EIRD). Finally, the IRs important for train scheduling were grouped into the third category, Planning and Scheduling Display (P&SD).

2.6 Display Prototype

Seventeen requirements applied to the P&SD (Table 1), which is the first display prototyped under this joint effort. Due to the traceability of the hCTA process, the information requirements can be linked to the situation awareness requirements (SAR) or display requirements (DL-DR), which can be associated with event flow diagrams and, in turn, the original scenario task overview. Therefore, our resulting list of requirements represents the complete set of data needed by a train operator to safely and efficiently manage scheduling for a passenger train from the departure station to the arrival station. With the hCTA process completed, the focus shifted to design of the P&SD.

Table 1. Information Requirements for P&SD

IR#	Description	Grouping
1	Current speed	SAD/P&SD
2	Goal speed	SAD/P&SD
3	Speed differential	SAD/P&SD
4	Current Traction/Friction lever position	SAD/P&SD
5	Current time	P&SD
6	Impact of event (system error, weather, etc.) on schedule	P&SD
7	Suggested speed profile	P&SD
8	Departure time	P&SD
9	Time to departure	P&SD
10	Potential impact of rail grade on speed	P&SD
11	Upcoming speed change indication	SAD/P&SD
12	Scheduling anomaly alert	P&SD
13	Train route with current location	SAD/P&SD
14	Next waypoint with scheduled arrival time	SAD/P&SD
15	Difference between predicted arrival and scheduled arrival at next waypoint	SAD/P&SD
16	New recommended speed profile with impact on schedule	P&SD
17	New recommended goal speed	P&SD

A well-designed interface, with the appropriate use of visualizations and display elements, can support decision-making [11] and minimize the cognitive complexity of a task [12]. Many design principles and usability heuristics were referenced to guide the development of the P&SD interface to ensure that the display could adequately support the train operator. The design principles used to guide the design included the *Principle of Information Need*, the *Principle of the Moving Part*, and the *Principle of Proximity-Compatibility* [13]. Usability heuristics, while not formal design principles, were also used to guide interface development. Many experienced design experts have devised lists of "best practices", including Nielsen [14], Tognazzini [15], and Schneiderman [16], and the heuristics of *Consistency*, *Recognition Over Recall*, and *Simplicity* were also applied to the design.

The resulting display included six main functional groups (Figure 4): 1) Title Bar summarizing high-level trip details, 2) Trip Planning Bar providing schedule updates and new proposed speed profiles, 3) Route Overview Bar providing an overview of the train route from departure location to arrival location, 4) Speed Profiles Bar visually depicting important speed profiles, including maximum acceptable speed, minimum acceptable speed, and suggested speed, 5) Terrain Profile Bar depicting terrain changes through the route, and 6) an Information Dashboard summarizing important data related to the trip, including scheduled time of arrival at the next waypoint, current speed, and current track grade.

With a display prototype complete, the next step is to evaluate the prototype through usability and performance assessments, including testing in a high-fidelity simulation environment. Long-term goals are to determine a technology transition path towards system integration.

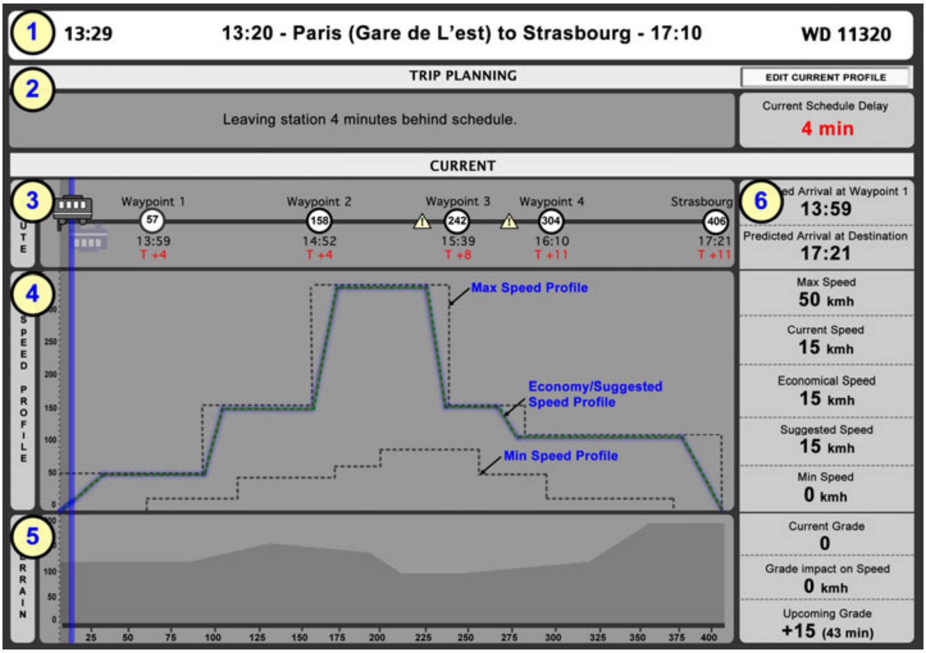


Fig. 4. Annotated Planning and Scheduling Display

3 Conclusion

This paper described the analysis of a complex work environment, train operation, with the ultimate goal of deriving requirements for a train schedule management interface. A Hybrid Cognitive Task Analysis (hCTA) was used to generate the information requirements for the engineer of a passenger train through departure to arrival. The hCTA included constructing a scenario task overview and converting it into event flow diagrams for each identified operational phase. Complex decisions were then analyzed using ladders, which allowed for the derivation of related display requirements. Finally situation awareness requirements were derived for the remaining operational tasks. In order to accurately depict the continuous monitoring tasks that frequently occur during train operation within the event flow diagrams, a new symbol was created, termed the *monitoring block*. The addition of this symbol extended the temporal bounds of traditional event flow diagrams, allowing for the depiction of simultaneous continuous-monitoring blocks.

The result of the hCTA included seventeen information requirements that were directly related to the scheduling task. These requirements were then used to guide the development of a Planning and Scheduling Display (P&SD) to be used within train cabs to assist operators with schedule management. With a display prototype complete, the next step is to assess the usability and performance aspect of the display.

References

1. Houpt, P.K., Bonanni, P.G., Chan, D.S., Chandra, R.S., Kalyanam, K., Sivasubramaniam, M., Brooks, J.D., McNally, C.W.: Optimal Control of Heavy-Haul Freight Trains to Save Fuel. In: 9th International Heavy Haul Association Conference, pp. 1033–1040. IHHA, Virginia Beach (2009)
2. Nehme, C.E., Scott, S.D., Cummings, M.L., Furusho, C.Y.: Generating Requirements for Futuristic Heterogeneous Unmanned Systems. In: HFES: 50th Annual Meeting of the Human Factors and Ergonomic Society, pp. 235–239. HFES, Santa Monica (2006)
3. Scott, S.D., Cummings, M.L.: Cognitive Task Analysis for the LCS Operator. Technical Report. Humans and Automation Laboratory (MIT), Cambridge (2006)
4. Carrigan, G.P.: The Design of an Intelligent Decision Support Tool for Submarine Commanders. Thesis. MIT, Cambridge (2009)
5. Tappan, J.M., Pitman, D.J., Abi Akar, C., Cummings, M.L.: Minimum Information Interface for Locomotive Operations (MIILO). Technical Report. Humans and Automation Laboratory (MIT), Cambridge (2010)
6. Shepherd, A.: Hierarchical Task Analysis. Taylor & Francis Inc., New York (2001)
7. Rasmussen, J., Pejtersen, A., Goodstein, L.: Cognitive Systems Engineering. John Wiley & Sons, Inc., New York (1994)
8. Rasmussen, J.: Skills, rules, and knowledge: Signals, signs, and symbols, and other distinctions in human performance model. IEEE T. Syst. Man Cyb. 13(3), 257–266 (1983)
9. Sheridan, T.B., Verplank, W.: Human and Computer Control of Undersea Teleoperators. Technical Report. MIT, Cambridge (1978)
10. Endsley, M.R.: Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors 37(1), 32–64 (1995)
11. Ware, C.: Information Visualization: Perception for Design. Morgan Kaufmann, San Francisco (2004)
12. Guerlain, S., Jamieson, G., Bullemer, P., Blair, R.: The MPC Elucidator: A case study in the design of representational aids. IEEE T. Syst. Man Cyb. 32(1), 25–40 (2002)
13. Tsang, P., Vidulich, M.A.: Principles and Practice of Aviation Psychology. Lawrence Erlbaum Associates, Mahwah (2003)
14. Nielsen, J.: Ten Usability Heuristics, `http://www.useit.com/papers/heuristic/heuristic_list.html`
15. Tognazzini, B.: First Principles of Interaction Design, `http://www.asktog.com/basics/firstPrinciples.html`
16. Schneiderman, B.: Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison-Wesley, Reading (1987)

Part V
Security and Safety

Application of Natural Language in Fire Spread Display

Yan Ge, Li Wang, and Xianghong Sun

State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, CAS
4A Datun Road, Chaoyang District, Beijing 100101, China
{gey,wangli,sunxh}@psych.ac.cn

Abstract. How to express fire spread efficiently and effectively to firefighters was an important issue. The present study aimed to investigate how to present the fire alarm information, focusing on whether natural language alarm presentation was better than the alarm list presentation. Objective method and subjective evaluation were used to compare the difference among different expressions. The results revealed that natural language was better than alarm list in described a fire spread situation, and the effect was more robust when spatial information was added. Traditional alarm list was more accuracy than other forms, but it cost more time to read and comprehension. So natural language with spatial information will be recommended to the future design of fire alarm system.

Keywords: natural language, alarm list, fire spread display, comprehension.

1 Introduction

Time is one of the most important issues in firefighting. Firefighters need to access fire information as quickly and accurately as possible. How to present fire spread efficiently and effectively to firefighters was an important issue for relative product design. In previous studies, we explored some interactive styles of how to present information in fire situation display systems [1, 2], but we still did not know the role of language structure in comprehension of fire spread display.

The production and reception modes of natural language are crucial issues in human-computer dialogue. Le Bigot et al. examined the effects of user production and user reception modes on natural language human-computer dialogue. They found the most efficient configuration for interacting in natural language would appear to be speech for production and system prompts in text, as this combination decreases the time on task while improving dialogue involvement [3]. We also found the similar results in an En Route display of Fire Information [4]. With advances in computer technology, text analysis allows researchers to reliably and quickly assess features of what people say as well as subtleties in their linguistic styles [5]. Text Analysis offers optional solutions for information extraction and natural language processing. People always display adaptive language behaviors during human computer interaction [6]. Based on these result, how to express the fire spread situation became an important question.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 365–373, 2011.

This study aimed to investigate how to present the fire alarm information, focusing on the effect of language structure on fire scenarios comprehension. We adopted an objective choice task and a subjective evaluation task to compare four different presentation forms: Alarm List, Integrated Alarm List, Natural Language and Natural Language with Spatial Information. The following questions were needed to answer: could natural language presentation provide enough information as alarm list, and help participants to obtain fire spread information accurately? Which form could be processed most quickly? Is it useful to add spatial information in natural language form?

2 Method

One factor with four levels design was used to determine the effect of different fire alarm presentations. Objective method and subjective evaluation were used to compare the four presentation forms. The main task of participant was to choose one correct fire scenario from four choices after reading/listening this fire scenario. The response accuracy, response time, times of switch between description and simulation, description viewing time and simulation viewing time were recorded.

2.1 Participant

12 firefighters from two fire brigade participated in this experiment, aged from 24 to 29. All reported normal or corrected vision.

2.2 Materials

4 presentation forms were used to express fire spread situations. The examples of them were giving in Table 1. Form 1 was a traditional alarm list, used as a baseline. It only included specific location and time of every activated smoke alarm. Form 2 was an integrated alarm list, which integrated information of smoke detectors in the same floor. Form 3 was a natural language alarm. Temporal and spatial information were integrated in to a whole sentence, like firefighters used to describe a new fire to their colleagues. Form 4 a natural language alarm with spatial information. The spatial information of the first detector was included in this form in order to investigate the value of supplementary spatial information.

In the learning part, 12 fire scenarios were used as materials in this experiment, including 8 one origin fires and 4 two origin fires. Each scenario was displayed in 4 forms as described above, so 48 presentations were prepared all together in this test. All 48 presentations were divided into four groups by Latin square method. Each group included 12 scenarios and every three scenarios used one presentation form in order to counterbalance the difficulty of 12 scenarios and their description forms. Each participant was arranged to choose one group material randomly to complete the test. The sequence of 12 scenarios was randomly presented.

Table 1. Examples of 4 forms used to express fire spread situations

Forms	Example
Form 1 Alarm list	14:22 ground floor, Break 026 14:23 ground floor, Corr 030 14:23 ground floor, Corr 030 14:23 ground floor, ELEC 028 14:25 Floor 1, ELEC 134 14:25 ground floor, Corr 030 14:25 ground floor, Corr 030 14:25 floor 1, Corr 130 14:25 floor 1, Corr 130
Form 2 Integrated Alarm list	At 14:22, first smoke detector in Break 026 of ground floor was activated. At 14:23, 3 additional detectors in ground floor were activated. At 14:25, smoke detector in Elec 134 of floor1 was activated, 2 additional detectors in ground floor were activated. At 14:26, 2 additional detectors in floor1 were activated
Form 3 Natural Language	At 14:22 smoke was first detected on the Ground Floor in Break Room 026. It spread along the corridor to Room 030. At 14:25, smoke spread to Floor 1 near Room 130.
Form 4 Natural Language with Spatial Information	At 14:22 smoke was first detected on the Ground Floor in Break Room 026.,which was at northwest of the building. It spread along the corridor to Room 030. At 14:25, smoke spread to Floor 1 near Room 130.

In the testing part, four simulated animations were produced for each scenario to test the comprehension of firefighter for each scenario. One was a correct simulation which matches the scenario description, while the other three were false simulations with little mistake. The precondition for setting false simulations was that the number of activated detectors in correct simulation was equal to that in false simulations. In false simulation 1, only the activated time of some smoke detectors was different from the correct simulation. In false simulation 2, the first alarm was in different location of the same floor. In false simulation 3, the first alarm was in the different floor. Four simulated animations were showed on the four monitors randomly during the experiment.

2.3 Procedure

Participants were comfortably seated in an instrumented room (See figure 1). Firstly, scenario descriptions were presented in two ways in the same time, the text information was presented on the first monitor and the audio information was presented by earphone. Then, the scenario descriptions disappeared and four simulated animations appeared on the four monitors separately. Participants' task was to make sure which simulation was matched with the description presented before. They were required to press the key marked "1", "2", "3", or "4" mapping to the monitor marked "1", "2", "3", or "4" to choose the correct answer. They could switch between the description and the simulation as they wish by pressing the "TAB" key during the task.

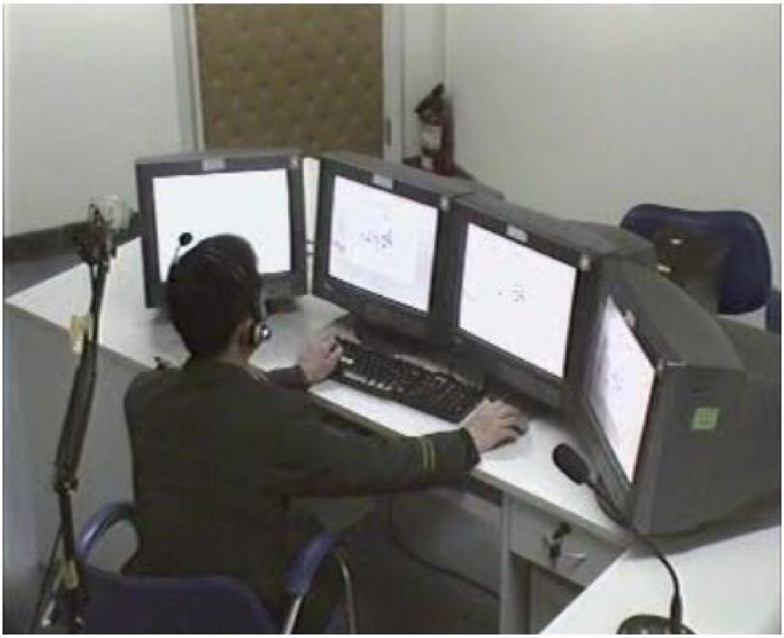


Fig. 1. Picture of experiment environment

There were three stages in the whole experiment: practice, test and interview. Before the formal test, participants have to practice using an additional scenario. Four expression forms were tested in this scenario. Participants could ask any question during this stage. In the test stage, 12 experimental scenarios were tested in four presentation forms. This stage was comprised of twelve trials. Test trials were presented in a new random order for each participant. After all the test trials were completed, participants were asked to rank the order of the four expression forms based on which one was easier to understand and to explain the reason.

3 Results

Six dependent variables were recorded for further analyze, including percentage of correct response, response time, description viewing time, simulation viewing time, switch times between description and simulation, and subjective ranking data. The data were analyzed with SPSS 13.0 using single-factorial analyses of variance with repeated measures. The effect of four expression forms (Alarm List, Integrated Alarm List, Natural Language and Natural Language with Spatial Information) was tested.

The statistical result of percentage of correct responses revealed a marginally significant effect of expression form, F (3, 33) = 2.73, MSE = .07, p < .1. Paired comparisons reflected that only the difference between alarm list (.86) and integrated alarm list (.56) was significant, t (11) = 3.52, p < .01. See Fig.2 for details.

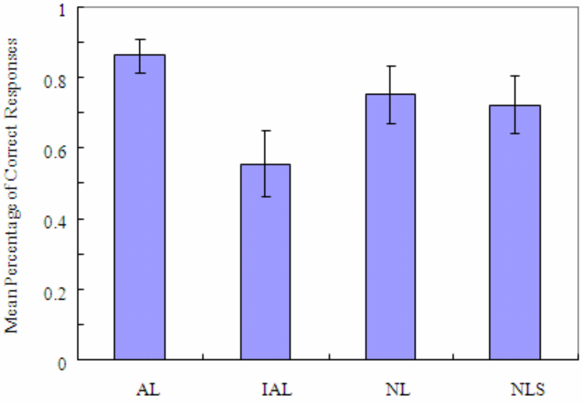


Fig. 2. Mean percentage of correct responses in each expression forms. Error bars represent standard errors. (AL: Alarm List; IAL: Integrated Alarm List; NL: Natural Language; NLS:Natural Language with Spatial Information).

The data of response time (duration between scenario description onsets and participants made choice) revealed no significant effect of expression forms, F (3, 33) = 1.38, MSE = 2915.75, p= .27. No significant result was detected in simulation viewing time, F (3, 33) = 1.30, MSE = 2426.67, p= .29.

The data of scenario description viewing time showed no significant effect of expression form, F (3, 33) = 1.73, MSE = 216.48, p= .18. Paired comparisons reflected that only the difference between alarm list (54.66 ms) and natural language with spatial information (41.13 ms) was significant, t (11) = 2.22, p < .05.See Fig. 3 for Means and SEs.

No significant effect of expression form was showed in time of switch revealed no significant effect, F (3, 33) = 2.09, MSE = 2.03, p= .12. But paired comparisons reflected that the difference between alarm list (5.35) and natural language with spatial information (4.25) was significant, t (11) = 2.30, p < .05, and the difference between integrated alarm list (5.59) and natural language with spatial information (4.25) was marginally significant, t (11) = 1.83, p = .09. See Fig. 4 for details.

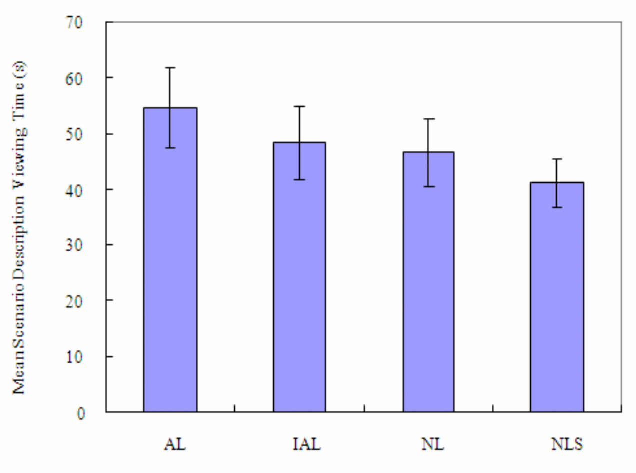


Fig. 3. Mean scenario description viewing time in each expression forms. Error bars represent standard errors.

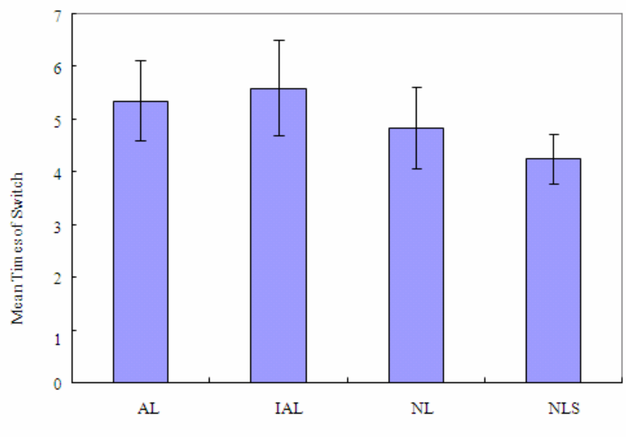


Fig. 4. Mean Times of Switch in each expression forms. Error bars represent standard errors.

The data of ranking scores were also analyzed. There was a marginally significant effect of expression forms, F (3, 33) = 2.90, MSE = 1.44, p= .05. Paired comparisons reflected that the difference between integrated alarm list (1.75) and natural language with spatial information (3.17) was significant, t (11) = 3.74, p < .01. The difference between integrated alarm list (1.75) and natural language (2.67) was marginally significant, t (11) = 2.20, p = .05. The difference between natural language(2.67) and natural language with spatial information (3.17) was also marginally significant, t (11) = 1.92, p = .082. See Fig. 5 for details.

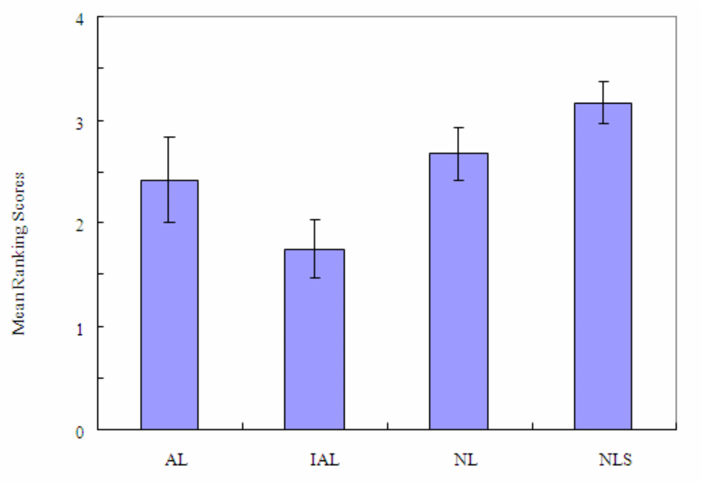


Fig. 5. Mean ranking scores in each expression forms. Error bars represent standard errors.

4 Discussion

The main findings showed that alarm list was better than integrated alarm list in accurate rate. The reason may be that integrated alarm list omitted some key information, so participants did not have enough information to make accurate responses. However, the difference between natural language and alarm list was not significant, neither did difference between natural language with spatial information and alarm list. Therefore it could be inferred that spatial and temporal integration in natural language provided enough information for participants to make correct responses.

Our focus was on whether natural language can help participants react quickly. The following speed indexes were considered: times of switch, scenario description view time, scenario simulation view time and response time. For response time, natural language with spatial information was processed quickly. Participants spent less viewing time when scenarios presented in natural language than in alarm list, especially in natural language with spatial information form. They spent the longest time to view the simulations when the scenarios were presented in integrated alarm list. Participants need to switch from description to simulation more often when scenarios were presented in integrated alarm list than in natural language with spatial information.

Four forms of fire scenario presentation were discussed respectively as below.

For the traditional alarm list, its accuracy of response was the highest one. But its response time was longer than natural language with or without spatial information. There was some trade-off in this description form. Further analyze showed its description viewing time was the longest one among four description forms. One reason for high accuracy and long description viewing time of traditional alarm list was it contained all information about a scenario, so its accuracy was high. But alarm list was too prolix to remember, so participants spent more time to view its description, switched much more times to choose correct simulation.

For the integrated alarm list, its accuracy of response was the lowest one. And its response time was the longest one among the four description forms. The subjective evaluation score of this form was also the lowest. It was the worst form among four expression forms. Compared with other forms, integrated alarm list missing some spatial and detail information in each floor. It was the hardest one to comprehension, so participant had not enough cues to judge which simulation was the correct one.

For natural language with or without spatial information of origin fire point, it was better than integrated alarm list in accuracy or response time. The response time, description viewing time and simulation viewing time of natural language with spatial information was shortest among four forms. So, for objective speed indexes, natural language was better than alarm list, and the effect was robust when natural language with spatial information. Therefore, it was necessary to add spatial information. We also considered the subjective index, ranking scores. Consistent with the objective indexes, there was a trend that natural language was better than alarm list, and the effect was more robust when scenarios were presented in natural language with spatial information form.

Many reasons could explain why participants preferred natural language than alarm list. First, there were too many detectors in alarm list. It was hard to remember when the number of detectors was beyond the capacity of human's working memory. Second, participants had to integrate information by themselves when reading alarm list. Third, alarm list was too prolix to comprehend, but natural language was more logical.

There were some problems we did not resolve in this study. Firstly, descriptions of scenarios were presented in text and audio simultaneously, but we cannot control whether they listen to the audio expression. Some participants reported that they were listening not so carefully, information about fire scenario was mainly obtained by reading text. This was maybe one reason for advantage of natural language was not so obviously. Secondly, the quantity of information in these four forms was not equal. Integrated alarm list missed some key information than other forms. Finally, we supposed quantity of traditional alarm list and natural language is equal, but the expression of alarm list is too long, participants had to spend more time to read and listen to this expression. How to control the length of different expression is needed to be discussed in the future.

5 Conclusion

Combined these objective indexes and subjective index together, it suggested that natural language is better than integrated alarm list, especially when fire origin point accompanied with spatial information. Traditional alarm list was more accuracy than other forms, but it cost more time to read and comprehension. For emergency scenes as fire fighting, the expression form which could cost fewer time and bring higher performance will be chose as the final description in real product to improve its quality. So natural language with spatial information will be recommended to be used in the future design of fire alarm system.

References

1. Sun, X., Plocher, T., Qu, W.: An empirical study on the smallest comfortable button/Icon size on touch screen. In: Aykin, N. (ed.) HCII 2007. LNCS, vol. 4559, pp. 615–621. Springer, Heidelberg (2007)
2. Qu, W., Sun, X.H.: Interactive Style of 3D Display of Buildings on Touch Screen. In: Harris, D. (ed.) HCII 2007 and EPCE 2007. LNCS (LNAI), vol. 4562, pp. 157–163. Springer, Heidelberg (2007)
3. Le Bigot, L., Rouet, J.F., Jamet, E.: Effects of speech- and text-based interaction modes in natural language human-computer dialogue. J. Hum. Factors 49, 1045–1053 (2007)
4. Qu, W., Sun, X., Plocher, T., Wang, L.: A Study of Information Retrieval of En Route Display of Fire Information on PDA. In: Jacko, J.A. (ed.) HCI International 2009. LNCS, vol. 5612, pp. 86–94. Springer, Heidelberg (2009)
5. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annu. Rev. Psychol. 54, 547–577 (2003)
6. Pearson, J., Hu, J., Branigan, H.P., Pickering, M.J., Nass, C.I.: Adaptive Language Behavirour in HCI: How Expectations and Beliefs about a System Affect Users' Word Choice. In: CHI 2006 Proceedings Conference on Human Factors in Computing Systems, pp. 1177–1180, Montréal, Québec, Canada (2006)

Differences between Students and Professionals While Using a GPS Based GIS in an Emergency Response Study

Rego Granlund[1], Helena Granlund[1,2], and Nils Dahlbäck[3]

[1] Santa Anna IT Research Institute, Linköping, Sweden
[2] Swedish Defence Research Agency, Linköping, Sweden
[3] Linköpings Universitet, Linköping, Sweden
rego.granlund@santaanna.se, helena.granlund@santaanna.se,
nilda@ida.liu.se

Abstract. This paper describes the results and differences between students and professionals who used a GPS based GIS as a collaborative tool in an experimental emergency response study. A total of 132 students, forming 22 groups and 108 professionals forming 18 groups were tested. Differences in both performance and behaviors between the groups have been identified. In the discussion we reflect on the importance to be aware of the participants' background and behaviors while selecting the participants in an experimental study.

Keywords: Experiment, Collaboration Support, Global Position Systems, Simulation, Emergency Management.

1 Introduction

This paper compares the results from a research project that includes two experiment studies. The experiment series was conducted with the same experimental method, but with different participant groups. In the first series, was the research idea tested on non-professional participants, a total of 132 university students. In the second series, 18 Swedish municipal crisis management teams, a total of 108 professionals were tested. The project explored the differences in work processes between teams that had access to GPS information in their digital map systems at command post level, compared to teams that only had paper maps (Johansson et al, 2010, Granlund et al, 2010, Granlund et al, 2011).

In Sweden many municipal organizations have made, or is about to make, investments in information and communications technologies, which provides crisis managers and rescue service personnel access to GPS information in all levels of command. The investments are done in order to enhance performance and control in response work.

GPS and digital maps are seen as support tools for crisis management that is understood to be more efficient with the introduction of these new technical supports. The reality is that they are tools that can distribute large amounts of information automatically to all users at all levels of management simultaneously. What originally

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 374–383, 2011.

was seen as an aid in the management work may have unforeseen consequences. The tools may change the requirements for managing and organizing emergency efforts.

1.1 Students vs Professionals

In many experimental studies, as is the case in much research, students are used as participants. There are of course some good reasons for this practice; one obvious one is the availability of large groups of participants, while using professional participants requires a laborious and time consuming process to recruit a large enough set of participants. And in addition, if working with rescue service professionals, there is always the risk that a real emergency occurs, forcing the data collection to be cancelled.

But using students as participants are not without problems, and this has been known for a long time. For an early example, see e.g. Smart's classical paper *Subject Selection Bias in Psychological Research* (Smart, 1966). Among the factors that Smart and his followers have pointed out where students differ from the general population are age, educational level, social background, where these in turn are known to co-wary with many variables potentially influencing performance in the tasks under study, e.g. intelligence, social skills etc.

Another factor which more recently has been brought to attention is that almost without exception student volunteers are exactly that, i.e. volunteers. If volunteering co-varies with factors potentially influencing the task under study, this limits the generalizability of the results obtained. Dahlbäck and Karsvall (2000) found for instance that the vast majority of participants were extrovert personalities, a factor which in turn is known to co-vary with communication patterns and risk-taking to mention a few.

All this calls for caution when using student volunteers as participants. But one can also note that all of the work mentioned above, as well as research reviewed in these papers, concern rather artificial lab experiments. It becomes therefore of interest to study whether student volunteers' performance differs from professionals in more ecologically valid settings, like for instance in microworld experiments. In this paper we will therefore compare both performance and communication patterns of students and professionals in exactly the same experimental situation.

2 Method

The method used for the two studies was controlled experiments in the microworld simulation C3Fire (Granlund, 2001; Johansson et al., 2007; Granlund et al., 2010). During the experiment series the participants experienced a set of scenarios, where an emergency response task was simulated. Half of the participating teams performed the task with a GPS support. Half of the participating teams were supported with paper maps.

2.1 C3Fire

The C3Fire microworld is a simulated environment where the system designers select important characteristics of the real system and create a small and well-controlled simulated system based on these characteristics. C3Fire often generates a dynamic forest fire fighting task and has been used extensively in previous research on network-based command and control (www.c3fire.org, Artman and Wearn, 1999; Granlund, 2001, 2004; Johansson et al., 2004), on effects concerning information support systems (Johansson et al., 2010; Granlund et al., 2010), on cultural differences in teamwork (Lindgren & Smith, 2006a, 2006b), and comes from a long tradition of microworld research on distributed decision making (Brehmer, 2005; Brehmer and Dörner, 1993).

The advantage of using a microworld is that the complex, dynamic and opaque characteristics that are generated by a proper microworld represents the cognitive tasks people encounter in real-life systems. Microworlds allows for the presentation of a number of different problems for the participants, rather than a single, well-defined task (Brehmer and Dörner, 1993; Dörner & Schaub, 1994; Granlund, 2001).

In this study the C3Fire system generates a task environment where participants can take on different roles, as commanders in a command post or fire fighting ground chiefs. The command post and the ground chiefs need to co operate to put out one or more forest fires. The simulation includes forest fire, houses, different kinds of vegetation, wind, and fire fighting units. The ground chiefs are responsible for the low level operation, such as the fire fighting, which is done in a short time frame. The command post works at a higher level and is responsible for coordinating the fire fighting forces and strategic thinking (Figure 1 and 2).

Computer based monitoring are integrated in the simulation and in all the information tools, used for the C3Fire environment. During a simulation the C3Fire environment creates a log with all events in the simulation and all computer mediated activities (Granlund, 2001).

2.2 Two Participant Groups

The two studies had different participant groups. In Study 1, conducted 2006, the participant group was Swedish university students. 22 student groups with six persons in each group were tested, giving a total of 132 participants (Johansson et al 2007, 2010).

In Study 2, conducted 2008-2009, the participant group was Swedish municipal crisis management members. Their average age was 49 years. 18 groups with six persons in each group were tested, giving a total of 108 crisis management members (Granlund et al, 2010; Granlund et al, 2011).

2.3 Experiment Design

Study 1 and Study 2 had the same between-group design with one factor: (a) crisis management teams using GPS, and (b) crisis management teams using paper maps (Figure 1 and 2). The difference between the two conditions is the type of support the participants obtain in terms of information visualization and data sources.

In each six person group, three participants worked as command post with one commanding officer and two liaison officers. Three participants worked as ground chiefs, controlling fire fighting units in the simulated environment.

In the GPS condition the commanding officer had access to a computer terminal equipped with a GPS that provided access to different digital map layers containing geographical information and exact positioning of the resources in the simulated world. The two liaison officers had computer terminals for communication with the ground chiefs in terms of e-mails (Figure 1).

In the paper map condition the commanding officer had no GPS access, but only a paper map of the simulated area. The two liaison officers had computer terminals for communication with the ground chiefs in terms of e-mails (Figure 2).

The ground chiefs in both conditions had access to a computer terminal with a single layer digital map and a communication tool that made it possible to communicate with the command post (Figure 1 and 2).

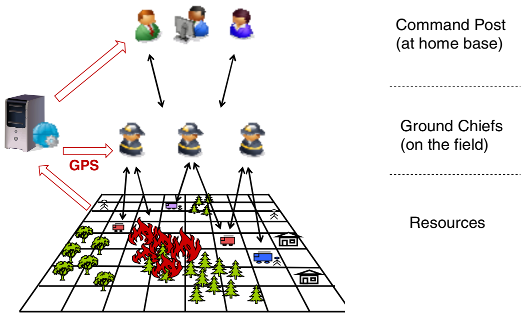


Fig. 1. The GPS condition

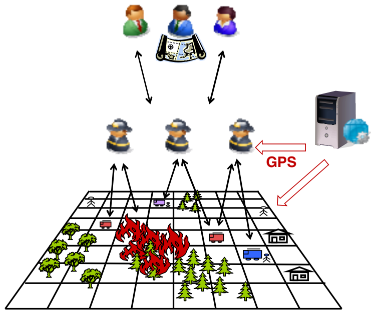


Fig. 2. The Paper Map condition

The resources were nine fire fighting units controlled by the ground chiefs. The task, the resources and the organization created the complexity of the experiment session.

2.4 Experiment Procedure

The experiment procedure included a customary training, then five regular simulation cycles. Each cycle include a 20 min simulation trial, 5 min of reflective questionnaires and 20 min of joint reflection, and after action review, where the group sees a fast recording of their actions during the simulation trial (Figure 3) (Granlund, 2008).

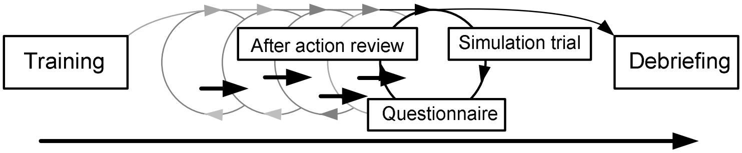


Fig. 3. The experiment procedure

3 Results

The results from the two studies, with students and professionals, are presented with respect to the two conditions GPS and Paper Map and with regards to performance and communication volume.

3.1 Performance

The main task for the participating teams was to control the forest fire and save houses. The measure of the success and performance of the team is a measure of the total amount of burned down area at the end of each simulation. A small amount on BurnedOutArea is preferable to a large.

Study 1. For students the results show, an over all significant difference, P=0,021 (N=132), between GPS and Paper Map over the five simulation trials (Johansson et al, 2010). The groups with a GPS support have lesser amount of BurnedOutArea, a better performance, than the groups who used Paper Maps (Figure 4).

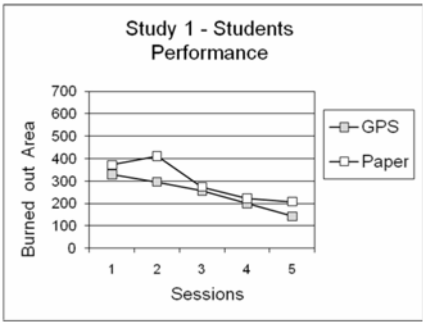


Fig. 4. Performance Student

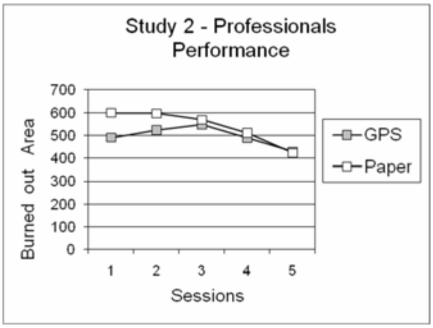


Fig. 5. Performance Professionals

Study 2. For professionals as participants there is no over all performance difference between GPS and Paper Map in the simulation trials and the trend of teams with GPS to have a smaller amount of BurnedOutArea is definitely broken in simulation session 5 (Figure 5). The professionals perform not as good as the students and have clearly inconsistent results compared to study 1.

One explanation for the results is that the command post of the participating groups is not uniform but has various compositions. The result above is divided into two subgroups. Teams where the command post consists solely of rescue service personnel, and subgroups in which the command post consist of a mix of municipal personnel (crisis management personnel), and rescue service personnel. The rescue service teams consist of a relatively homogeneous set of personnel with joint training and experience, and with a professional experience in managing crisis events. The mixed teams consist of a diverse group with a variety of training and professional experience, where one fraction of the participants is used to manage crisis events, while the other is familiar with management under normal conditions.

The result then shows that rescue personnel behave as expected from study 1 (Figure 6 and 7). But mix of municipal personnel and rescue personnel does not follow the patterns from study 1. The support tool did not help the mixed groups (Granlund et al. 2011).

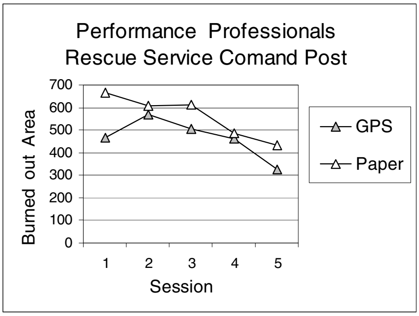


Fig. 6. Performance Rescue Service

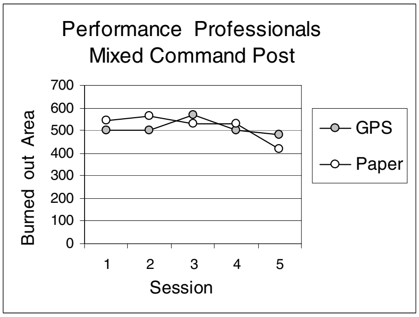


Fig. 7. Performance Mixed Teams

Summary Performance. Students have an over all significant difference in performance, where teams using GPS performs better than teams using Paper Maps. Professionals do not show any significant over all performance result between teams using GPS and teams using Paper Maps. When dividing the professional to sub groups, the rescue personnel behave as expected but not the mixed teams.

3.2 Communication Volume

This section present the volume of communication as amount of send messages between command posts and ground chiefs.

Study 1. The results for students show a difference in the patterns of communication between ground chiefs and command post in the GPS condition and the paper map condition in total over the five sessions (Figure 8). In the fifth simulation trial the command post of the GPS and paper map conditions send equally many text messages as the ground chiefs of the paper map condition, on average 80 messages. The ground chiefs of the GPS condition sent lesser than half of the amount, on average 30 text messages. This means that the ground chiefs in the GPS condition are relieved of workload when it comes to amount of sent messages, not the command post.

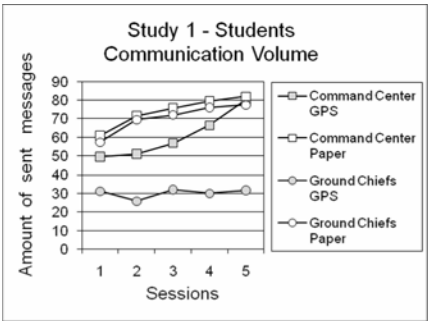


Fig. 8. Communication Students

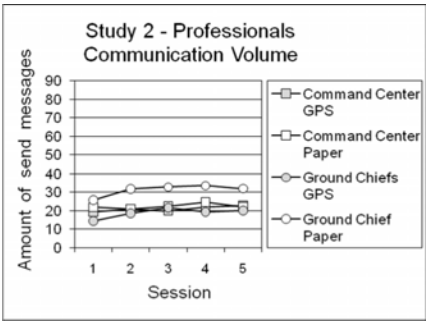


Fig. 9. Communication Professionals

Study 2. The crisis management personnel in study 2 show the same trend in the results for the amount of send messages (Figure 9). In the fifth session as many e-mails, about 22, are sent from the command post of teams that have access to a GPS support as from those who perform the task by using a paper map. This means that the command post is not relieved from work due to the GPS-based management support, with regards to the volume of communication. The second is that the ground chiefs in the GPS condition sends significant, t(16) = 3.13, p <.006, less messages than ground chiefs of the paper map condition. This means that it is the ground chiefs in the GPS condition that are relived of work load.

Summary Communication. Firstly there is a significant difference in amount of send messages between command posts in study 1 and study 2, t(78)=4.25, p<0.0001. The command posts in study 1 sent on an average 80 messages. The command posts in study 2 sent about 22. That is a significant difference. The ground chiefs in study 1 follow the same trend of sending more messages than ground chiefs in study 2.

By analyzing the amount of send messages from each entity, command post or ground chiefs, a picture of the teams' workload emerge. Study 1 and 2 shows two trends regarding communication workload for teams that use GPS. The first is that the command post is not relieved from work due to the GPS support. The second is that the ground chiefs operating on the simulated field are relieved from workload when their command post is supported by a GPS.

4 Discussion

The two participant groups, university students and crisis management professionals, differ in age, educational level and experiential background. The students are used to be evaluated and observed. Performing well during assessment is their means to show knowledge. The simulation system is far from important in their world of examinations. The professionals are not used to cameras, full-logging systems or researchers funded by the Swedish civil contingencies agency. Their way to demonstrate skills are based on good outcomes during response, not while being observed. Their management abilities are what they want to demonstrate.

4.1 Performance Students vs Professionals

An observation when comparing the performance measure in results from study 1 and study 2 is that the students does perform better in terms of BurnedOutArea than the professionals independent on if they were using GPS or Paper Maps (Figure 4 and 5).

The performance situation here is slightly different for the two groups. Students have none or a very limited experience of crisis management, paper maps or GPS. The students appreciated the game, appreciated to beat the game and found satisfaction in their group performance.

The professionals have a solid experience base to use when managing these kinds of efforts, especially when the means are paper maps. Their experience, skills and methods for crisis management make them connect the game to command and control aspects as well as crisis management in general. Situations that occur during simulations are close enough to their rough everyday responsibilities during response. They use the time for team and individual training. They appreciate the game as a means to find winning concepts from a response perspective.

Three possible explanations for the students over all better performance are computer game, professional experience and team composition. **1) Computer game:** The students have more experience from playing computer games than the professionals. A larger set of students did see the problem as a computer game task that they should solve. Win the game was often the goal. **2) Professional experience:** The professionals did behave as professionals when solving the task. They knew that their behavior was observed and analyzed in a research project. They used the strategies that they were use to do in real life, like not using all the resources directly, discuss before acting, etc. They did see the simulated task's similarities to a real situation. They did not try to win the game as a game; they tried to do what they should have done in a real situation. They could relate the events in the simulation to their profession. Hence they practiced and reflected on strategies that they utilize in real life. **3) Team Composition:** When the results were divided into the two team compositions interesting results appeared. Both the groups were professionals that work with crisis situations in their professions. Still they have different experiences and work tasks. Two main observations were that the mixed groups did not know each other as the rescue service teams did. A lot of task for the mixed group was to learn to work together as a team. Some of the mixed groups did see the day as a team training day. Another observation was that the command style did change depending on the profession.

An important observation is that if we compare the performance when using paper maps or GPS support it shows that both the students and the rescue service personnel performs better with GPS support.

4.2 Communication Students vs Professionals

Students communicate by far more via the text messages than professionals. The professionals are trained to communicate effectively, without adding to the confusion of the emergency situation, without taking precious time, and probably with a predisposition of consensus. The students were selected as volunteers and are likely the most extrovert partition of university students in general. The professionals were chosen because of their ability as crisis managers, being extrovert is far from a required feature in times of crises.

In relation to the GPS both students and professionals have the same results; the command post is not relieved from work due to the GPS support and it is the ground chiefs operating on the simulated field that are relieved from workload.

4.3 The Complexity of Measuring Performance

Measuring performance is complex. In study 1 and 2 the participants' task was to close out the forest fires. The basic performance measure should then be to count the burned out area as shown above. But the simulated world also contained houses etc, they too could be analyzed as a performance measure. The result of amount of BurnedOutHouses shows the same pattern. The students preformed better. It is not because the students considered the houses; they did not even mention them in their text messages. It is a consequence of their one goal, to extinguish fire effectively, out of a game point of view. The fire newer even reached the houses.

The professionals on the other hand, considered and mentioned the houses. They prioritized among them and agreed on rules for what to save and what to let burn. All on an experience based logic in order to practice what they need to know in real life. They fought for the houses they prioritized, and often, especially with the support of GPS, saved them, but not always and not as effective as the students, who did not even recognize the goal to save houses.

The question then is; what is performance, who performed best students or professionals?

References

1. Artman, H., Waern, Y.: Distributed cognition in an emergency Co-ordination Center. Cognition, Technology & Work 1, 237–246 (1999)
2. Brehmer, B., Dörner, D.: Experiments With Computer-Simulated Microworlds: Escaping Both the Narrow Straits of the Laboratory and the Deep Blue Sea of the Field Study. Computers in Human Behaviour 9, 171–184 (1993)
3. Brehmer, B.: Microworlds and the circular relation between people and their environment. Theoretical Issues in Ergonomics Science 6(1), 73–93 (2005)
4. Dahlbäck, N., Karsvall, A.: Personality Bias in Volunteer Based User Studies? In: Proceedings of HCI 2000, vol. 2, pp. 49–50 (2000)
5. Dörner, D., Schaub, H.: Errors in Planning and Decision Making and the Nature of the Human Information Processing. Applied Psychology: An international review, Special Issue on Human Error, 433–453 (1994)
6. Granlund, H.: Experiential Learning in computer based simulation training - Experiences from research on team decision making. In: proceedings of 2008 International Conference on Information Technology in Education within CSSE, Wuhan, China (2008)
7. Granlund, R.: Web-based micro-world simulation for emergency management training. In: In Future Generation Computer systems, vol. 17, pp. 561–572. Elsevier, Amsterdam (2001); (best papers form the conference Websim 1999)

8. Granlund, R.: Monitoring experiences from command and control research with the C3Fire microworld. Journal Cognition, Technology and Work 5(3), 183–190 (2004), ISSN 1435-5558
9. Granlund, R., Granlund, H., Johansson, B., Dahlbäck, N.: The Effect of a Geographical Information System on Communication in Professional Emergency Response Organizations. In: Proceedings of ISCRAM 2010, 7th International Conference on Information Systems for Crisis Response and Management (2010)
10. Granlund, R., Granlund, H.: Using Simulations to Study the Impact of GPS Information in Crisis Response Organizations. Submitted to ISCRAM 2011, 8th International Conference on Information Systems for Crisis Response and Management, Lisbon, Portugal, May 8-11 (2011)
11. Johansson, B., Trnka, J., Granlund, R.: the Effect of Geographical Information Systems on a Collaborative Command and Control Task. In: Van de Walle, B., Burghardt, P., Nieuwenhuis, K. (eds.) Proceedings of ISCRAM 2007, pp. 191–201. Delft, the Netherlands (2007)
12. Johansson, B., Trnka, J., Granlund, R., Götmar, A.: The Effect of a Geographical Information System on Performance and Communication of a Command and Control Organization. The International Journal of Human-Computer Interaction. Special issue on Naturalistic Decision Making with Computers. 26(2&3), 228–246 (2010)
13. Lindgren, I., Smith, K.: Using microworlds to understand cultural influences on distributed collaborative decision making in C2 settings. In: Proceedings of the 11th The International Command and Control Research and Technology Symposium (ICCRTS), Cambridge, UK (2006); (Awarded the Willard S. Vaughan, Jr. Best Student Paper Award)
14. Lindgren, I., Smith, K.: National patterns of teamwork during an emergency management simulation. In: Proc. 50th Annual Meeting of the Human Factors and Ergonomics Society, San Francisco, CA (2006)
15. Smart, R.: Subject Selection Bias in Psychological Research. Canadian Psychologist 7a, 115–121 (1966)

Adversarial Behavior in Complex Adaptive Systems: An Overview of ICST's Research on Competitive Adaptation in Militant Networks*

John Horgan[1], Michael Kenney[1], Mia Bloom[1], Cale Horne[1], Kurt Braddock[1], Peter Vining[1], Nicole Zinni[1], Kathleen Carley[2], and Michael Bigrigg[2]

[1] International Center for the Study of Terrorism, Pennsylvania State University, 326 Pond Lab,
University Park, PA, 16802
[2] Center for Computational Analysis of Social and Organizational Systems (CASOS),
Carnegie Mellon University, 4212 Wean Hall, Pittsburg, PA 15213
{jgh11,mck14,mub27,cdh14,khb125,pbv5001,ncz5004}@psu.edu,
{kathleen.carley,bigrigg}@cs.cmu.edu

Abstract. There is widespread agreement among scholars and practitioners that terrorism scholarship suffers from a lack of primary-source field research [1]. The absence of solid ethnographic research has yielded studies that suffer from a lack of rigorous analysis and often result in opinion masquerading as analysis. This dearth of field work stems in part from a failure to integrate ethnographic research into computational modeling efforts. The project outlined in this paper seeks to redress this deficiency by combining the strengths of ethnographic field research with sophisticated computational models of individual and group behavior. Specifically, we analyze data from interview transcripts, news reports, and other open sources concerning the militant activist group Al-Muhajiroun and the terrorist groups Provisional Irish Republican Army (PIRA) and Revolutionary Armed Forces of Colombia (FARC). Using *competitive adaptation* as a comparative organizational framework, this project focuses on the process by which adversaries learn from each other in complex adaptive systems and tailor their activities to achieve their organizational goals in light of their opponents' action.

Keywords: Al-Muhajiroun, competitive adaptation, network analysis.

1 Introduction

A growing number of scholars and practitioners recognize the value of mixed-methods and interdisciplinary approaches to studying non-state actors that engage in political violence themselves or support the use of political violence by like-minded groups.

* This research is supported by a grant from the Office of Naval Research (ONR), U.S. Department of the Navy, No. N00014-09-1-0667. Any opinions, findings, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research. We thank Devan Zeger for extensive help in the preparation of this manuscript.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 384–393, 2011.

Furthermore, there is a growing consensus that the terrorism and counterterrorism literatures suffer from a lack of primary-source field research. The vast majority of terrorism scholarship is based on secondary sources, contributing to a systematic bias based in data availability. Naïve and impractical policy prescriptions are often the result. This paper addresses these issues and provides an interdisciplinary framework from which to study the behavior of non-state militant groups that either carry out acts of political violence themselves or support the use of political violence by others. Building on the concept of competitive adaptation, we investigate how these actors learn and adapt when interacting with governments, civilians and other militant groups, as well as how they alter their subsequent behavior. Some guiding questions behind our work include:

- How do militant and government networks learn from and adapt to one another over time?
- How do major events (terrorist attacks, counterterrorism operations, death of a group leader, government policy shifts) affect the structure of militant networks?
- Does the structure of militant networks differ in models based on open-source information (news account, court transcripts, official statements) versus private information (interviews with de-radicalized or disengaged militants)?

Our approach combines the analytical richness of ethnographic research with computational modeling to provide a meso-level model of militant networks that function in complex-adaptive systems. Specifically, we use data from original interviews with group members together with news reports, public statements and other open source documents to study the competitive adaptation of three groups across four countries. The cases studied – Al-Muhajiroun, the Provisional IRA and FARC – are non-random, selected on the basis of convenience of access, team expertise and dissimilarity to one another, for the purpose of developing a generalizable, multi-disciplinary approach to the study of such opaque organizations. Preliminary analysis presented below involves Al-Muhajiroun, a former militant activist group based in the United Kingdom, banned by British authorities in 2010. Next, we present the theoretical framework of competitive adaptation and its usefulness in the investigation of our research questions. The subsequent discussion of Al-Muhajiroun details methods of data collection and analysis.

1.1 Organizational Learning and Competitive Adaptation in Militant Groups

There are few studies that specifically examine learning in nonconventional, covert, illegal or violent groups such as militant or terrorist organizations. Nevertheless, some recent work offers useful insights into how militant and terrorist organizations learn and adapt within the adversarial environments in which they operate. A team of RAND researchers, led by Jackson [2], examined organizational learning in several violent militant groups, including the Provisional Irish Republican Army, Aum Shinrikyo, Jemaah Islamiyah, Hezbollah, and the radical environmentalist Animal Liberation Front and Earth Liberation Front. A separate study by Hamm [3] [4] draws on court documents contained in the *American Terrorism Study* database and the criminological literature on social learning to explore how some violent political extremists acquire the skills to perform their tradecraft. While these studies offer insights into how numerous militant groups train their members and develop certain technological innovations, they do not systematically examine the *internal* processes of group learning and interpretation, as experienced by militants themselves. Moreover, these studies do not take into account the broader competitive environments in which militant groups operate.

Drawing on organizational and complexity theory, Kenney [5] describes how organizational knowledge is leveraged by competing networks that interact in complex adaptive environments. Kenney dubs this process *competitive adaptation*, which explains how organizational learning occurs within an environment that is typically (though not always) characterized by hostility and multiple actors pursuing opposing goals. A network-based theoretical approach to the study of militant groups allows both effective modeling of the internal organizational dynamics of militant groups, in addition to broader strategic interactions between militant groups and governments. Competitive adaptation is thus the framework from which we approach our study of militant groups, and we employ ethnographically-based network analysis as our primary tool when modeling this framework. This approach, along with findings in the organization theory literature, provides us with guidance in developing our research and expectations.

1.2 Comparing the Social and Organizational Qualities of Militant Networks

Research on Al-Muhajiroun [6], the FARC [5] and Provisional IRA [7] provide qualitative descriptions of these groups' structures and how those structures may account for group behavior. We seek to expand on these findings using ethnographic data to compare and contrast these groups from both a qualitative perspective and using quantitative metrics of social network analysis. Measures of closeness, for example, allow for quantitative comparison of the distances between actors within networks, whereas measures of centrality enable us to compare the density of militant social networks [8]. Other metrics permit us to compare additional network attributes such as hierarchies and emerging leadership, thus enabling us to study the evolution of organizational attributes over time based on the analysis of primary and secondary ethnographic data. In the competitive adaptation framework, we expect that:

H1: Social and organizational network properties, such as centralization, hierarchy, and density of network ties, vary across different militant movements, and within movements across time in response to changes in their complex-adaptive environments.

Social network metrics calculated from the ethnographic data described above will allow us to arrive at substantive conclusions when comparing the structures, decision-making, learning and adaptation of militant and terrorist groups.

1.3 Examining Networks of Locations, Events, Knowledge, Resources and Tasks

When studying organizational learning and competitive adaptation in militant and terrorist networks, we seek to examine militant networks in their entirety, beyond the social context. Specifically, we recognize that in order to understand group dynamics, learning, evolution, decision-making and emergent behavior, it is necessary not only to examine the roles and relationships of individual agents and groups within organizations, but also how those agents relate to locations in space, as well as the knowledge and resources leveraged by agents within organizations, in order to fulfill group tasks. Carley [9] writes that an organization can be described as an "ecology of networks" that continually evolves as agents within the organization learn, move and interact. A network of social roles within an organization might appear very different from a network of knowledge and expertise, which in turn might be very different from the network of resources or geographic proximity. Kenney's [5] work on competitive adaptation similarly emphasizes the importance of organizational properties beyond those associated with individual human agents, arguing that the flow of knowledge, routines and artifacts within organizations is as important as the flow of personnel. We conceptualize militant networks consistently with these arguments and expect that:

H2: Militant networks 'learn' when their participants receive information about their activities, process this information through knowledge-based artifacts, and apply the information to their practices and activities.

The case study presented below does not definitively answer the project's two broad hypotheses, but offers an example of the detailed, preliminary analysis needed in preparation for meaningful hypothesis testing.

2 Case Study: The Evolution of Al-Muhajiroun

Following the development of our interdisciplinary framework, the research team began to collect primary and secondary data related to the first militant group we chose to study: Al-Muhajiroun (henceforth referred to as AM). While AM is not a terrorist group, the group did support the use of political violence overseas (outside of Britain) in what it maintained was a defensive response to the aggressive foreign policies of Western states, including Britain and the United States, before being formally banned by British authorities in 2010. Moreover, in recent years numerous militants that were affiliated, in varying degrees, with AM or one of its splinter groups have been convicted of what Simcox et. al. describe as "Islamism related offenses" involving political violence [11].

Founded in 1996 by Omar Bakri Mohammed and officially disbanded for several years in 2004, former AM members continue to operate in the United Kingdom under several splinter groups included in our study. Work by Wiktorowicz [6] indicates that AM not only exhibits adaptive behavior as it interacts with British authorities, but that the relatively liberal political and social environments of the United Kingdom often condition these interactions, providing both advantages and disadvantages to each side. As an example, Wiktorowicz shows how freedom of the press in the United Kingdom has been a double-edged sword for AM, both assisting the group in publicizing its ideas to potential recruits, but also resulting in broad ostracism of the group that has had direct negative ramifications on its operations. These negative ramifications have included the groups' banning from the use of public venues and increased police scrutiny of group activities, resulting in arrests and the costly loss of charitable organization status in the United Kingdom (p. 126).

In response to the negative ramifications of publicity, AM has adapted a strategy of organizational proliferation, diversification and obfuscation in order to continue spreading the group's ideology and connect with potential recruits without suffering the costs that are now associated with the AM label, and without risking organizational death in the event of a police crackdown (p. 124). Indeed, research by Raymond [12] shows that despite such a crackdown in 2004 which resulted in the organization's abolishment, AM members continue to operate in the United Kingdom and abroad under a variety of alternative platforms, fronts and splinter groups. Recent work by Kenney [13] discusses this adaptive dynamic between the British government and AM in depth. Following the disbanding of AM in 2004, the group's leadership established two new groups called Al Ghurabaa (The Strangers) and the Saved (or Saviour) Sect, both of which attracted many of the same members as AM. Furthermore, when these splinter groups faced the threat of disbandment, former AM leadership created the 'Ahlus Sunnah wal Jamaah,' an invitation-only Internet discussion forum (p. 124). More recently, former AM members and affiliates have created several new platforms or groups to facilitate their ongoing activism, including Muslims against Crusades and Supporters of Sunnah.

Whereas AM's organizational expansion is a clear example of adaptive behavior within the competitive environment in which it operates, it is also equally interesting that militant organizations such as AM often fail to adapt, or learn the wrong lessons, despite their experiences. Kenney [14] explains that militant groups might fail to adapt within their environments not only due to simple mistakes and human error, but potentially due to the underlying structures, ideologies or rules guiding an organization's behavior. The religious underpinnings of AM clearly condition the incentive structure of individual AM members and leaders, influencing how they adapt and fail to do so [6]. Kenney [14] notes that AM's religiosity has led its leadership to occasionally leave important tactical decisions, such as behavior that may result in imprisonment, to "God's predetermined fate for them." Thus from these examples of both adaptive and non-adaptive behaviors, AM is an interesting and relevant case study of how a militant group evolves and adapts (or has failed to evolve and adapt) in a Western democracy.

2.1 Collection of Primary and Secondary Source Data for Network Analysis

When collecting data for our analysis of AM, we have diversified among a broad range of primary and secondary sources. These sources included newspaper articles, interview and court transcripts, press releases and other group statements, as well as original interviews conducted with current and former AM members, members of front, successor and splinter groups, and British authorities.

From these primary and secondary sources, we constructed a thesaurus of known AM members (agents). In addition, we created thesauri of AM front and splinter groups (of which there were many), events, locations, resources, tasks and knowledge. These data are then processed using the text analysis program AutoMap, and semantic and network analyses are conducted using the dynamic network analysis (DNA) program Organizational Risk Analyzer (ORA) [15].

The preliminary findings presented below are based on 1,079 newspaper articles published from 1996 to 2009 that deal with AM primarily. The articles were collected using Lexis Nexis. Duplicate articles and those not primarily concerned with AM were excluded. The data are presented dynamically, divided into three time periods segmented by major events in the history of AM, its members and associated organizations. The first time period (Network A) runs from the group's founding in 1996 through October 4, 2004 when AM officially disbands. The second period (Network B) begins with the July 7, 2005 attacks on the London metro (an event which sparked Omar Bakri's flight to Lebanon, where he still resides) to July 16, 2006, the day before the UK government officially banned AM's successor organizations, Al Ghurabaa and the Saved Sect. The final period (Network C) runs from this July 17, 2006 ban through the end of 2009. This analysis is limited to AM members and associated Islamists (i.e., the 'red team') – a total of 364 individuals spanning the 14-year period in view.

2.2 Preliminary Findings for Al-Muhajiroun

Table 1 summarizes the changing relationship of Omar Bakri, AM's founder, to others in the AM network. The numbers in each column indicate cumulative totals (i.e., two degrees of separation gives the cumulative total of rows 1 and 2, and so forth). The totals given in the bottom row indicate the total number of 'red team' agents detected in the network during the period in view.

Several characteristics of Bakri's connectedness to the network merit discussion. First, during any time period, if an agent is not connected to Bakri, that agent is not connected to *anyone* in the network; that is, that node is an 'isolate.' The data used do not support a relationship between these individuals and the AM network, though ethnographic data may augment this conclusion. Second, at any point in time, every agent in the network who is connected to Bakri is connected by no more than four degrees of separation. Third, Network A is a superset of Networks B and C. In other words, no new agents connect to Bakri subsequent to the first time period. However, consistent with *H1*, Bakri's connectedness to other agents in the AM network changes dramatically across the three time periods in view.

Table 1. Omar Bakri: Sphere of Influence

Degree of Separation	Network A	Network B	Network C
1	19	13	7
2	65	32	14
3	83	41	18
4	84	42	23
Total	186	80	91

Figure 1 depicts this changing relationship. As might be expected, Bakri's direct connections to other agents in the network declines significantly following his exile to Lebanon, shortly after the 7/7 attacks on the London metro. In Network A, Bakri is connected to 83 of 84 red team agents by no more than three degrees of separation (only Adel Abdel Bary, indicted in the United States in connection to the 1998 U.S. embassy bombings, is separated from Bakri by four degrees). From Network A to

Network B, where the July 7, 2005 cutoff approximates Bakri's removal to Lebanon, his total connections within the network drops by exactly half, from 84 to 42.

Also, while everyone remains connected to Bakri within four degrees, the proximity of these connections declines following Bakri's move to Lebanon and the ban on AM's successor groups. In Network A, 77 percent of agents (65 individuals) are connected to Bakri by no more than two degrees of separation; in Network C this number drops to 60 percent (14 individuals). The ability to influence other agents does not normally extend beyond the second degree, suggesting that Bakri's influence over AM members in Network C is only one-fourth of his influence in Network A.

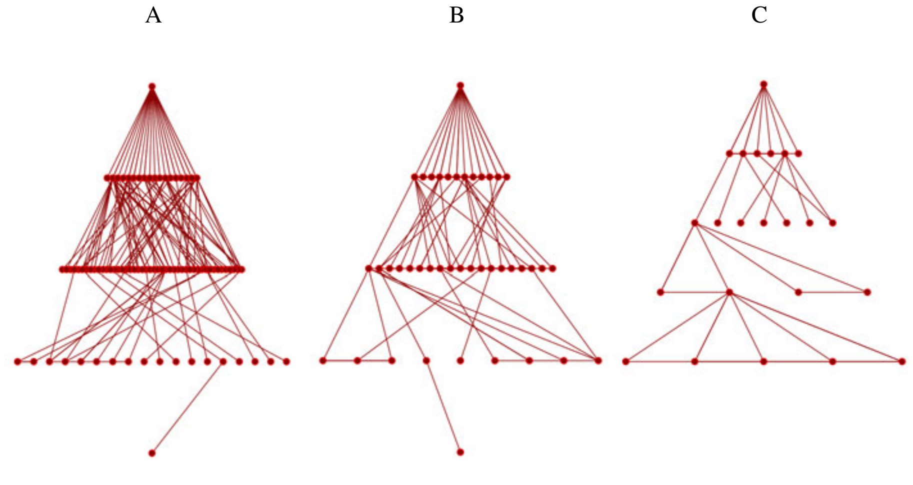


Fig. 1. Omar Bakri's changing sphere of influence

Further, in Network C, Bakri's connectedness beyond two degrees is tenuous: Only Osama bin Laden connects Bakri to the third degree, and only Mohammed Omran connects Bakri to the fourth degree. The bin Laden connection is based on circumstantial evidence at best. This shift suggests a gradual distancing between Bakri and the network, which is consistent with ethnographic evidence regarding the emergence of new leaders in the network, the proliferation of front and splinter groups in the network, and Bakri's apparent shift of roles into a position of symbolic leadership of the group from a distance. Future analysis of Bakri's changing role should examine changes over time in his access to resources for group mobilization. The extent to which Bakri's access to such resources improves or diminishes, combined with the access of his associates within two degrees, will provide significant information about the extent to which the AM network has been disrupted by 'blue team' efforts or has become a 'leaner, meaner' organizational machine (*H2*).

Other measures of leadership allow a comparison of Bakri's role in the network to the roles of other network elites. Agents in the network may score high along one or more dimension of leadership without ever holding a position of formal leadership in the organization. For example, Table 2 rank orders AM agents in terms of *betweenness centrality* across the three time periods, and Bakri ranks behind five to nine other agents in the network during any given period. Betweenness centrality measures the extent to which a given node (i.e., an agent) constitutes the most efficient path between other nodes in the network. In other words, this metric assesses which agents reside in the most 'best paths' between other agents, suggesting that individuals ranking high in this metric are likely to serve as brokers or gatekeepers between different subgroups within the network. While ethnographic research suggests that the presence of Osama bin Laden high in these rankings is likely an artifact of the data, the high betweenness centrality of other, lower profile agents merits discussion.

Abu Hamza al-Masri, for example, was a close associate of Bakri and leader of the Supporters of Shariah group, which frequently collaborated with AM in its UK activities. Anjem Choudary, Abu Izzadeen, Abu Uzair, and Abdul Rahman Saleem were all long-standing students of Bakri's that played different roles in AM and its splinter groups over the years. For example, Abdul Rahman Saleem served as AM's spokesman in Pakistan, and organized the movement of British Muslims into Pakistan and from there into the insurgency in Afghanistan. Once back in the UK, Saleem acknowledged that he received military training in Afghanistan and Pakistan, and sought to recruit other young Muslims to do the same.

Table 2. Betweenness Centrality

Rank	Network A		Network B		Network C	
1	Abdul Saleem	.042	Abdul Saleem	.078	Abu Izzadeen	.013
2	Osama bin Laden	.037	Abu Hamza	.060	Osama bin Laden	.012
3	Abu Hamza	.035	Abu Izzadeen	.051	Anjem Choudary	.011
4	Hassan Butt	.020	Osama bin Laden	.039	Abu Hamza	.009
5	Saladhuddin Amin	.013	Saladhuddin Amin	.027	Mohammed Omran	.009
6	Abu Qatada	.011	Omar Bakri	.020	Omar Khyam	.008
7	Omar Sharif	.011	Omar Sharif	.018	Abdul Saleem	.007
8	Waheed Mahmood	.010	Abu Uzair	.018	Abu Qatada	.007
9	Mohammed Omran	.010	Hassan Butt	.016	Omar Bakri	.005
10	Omar Bakri	.009	Anjem Choudary	.011	Mohammed Babar	.005

Eigenvector centrality offers a very different measure of agents' elite status within a network. This measure calculates the network's principal eigenvector, meaning that a given node is considered central to the network to the extent that its neighbors are central. Well-connected agents connected to other well-connected agents score high on this metric, while the formula discounts nodes possessing many connections, as well as accounting for the fact that most nodes will have some connections. Notably, Omar Bakri's eigenvector centrality ranking actually improves in Networks B and C, subsequent to his flight to Lebanon. While Bakri's direct connections to the network suffers while in exile, his continued association with the AM leadership (i.e., other well-connected nodes) elevates his centrality to the network in this measure.

Others, such as Abdul Kahar Kalam, Omar Sharif and Richard Reid, appear to score high along this measure as a function of their involvement in specific, high-profile terrorist plots. Such agents may be part of tight cliques, meaning they are highly connected to others in the clique while they do not encounter discounts for connections to many nodes in the network, which they do not possess. This is certainly the case with Ezzit Raad, arrested for his role in a terrorist plot in Australia, and Younis al Hayyari, an al-Qaeda affiliate shot dead in Saudi Arabia, both in 2005.

Table 3. Eigenvector Centrality

Rank	Network A		Network B		Network C	
1	Abdul Kahar Kalam	1	Abdul Kahar Kalam	1	Anjem Choudary	1
2	Omar Sharif	1	Omar Bakri	1	Abdul Saleem	1
3	Abdul Karim	1	Richard Reid	1	Willie Brigitte	1
4	Younis al Hayyari	1	Abdul Karim	1	Omar Bakri	.960
5	Abu Obeida	1	Younis al Hayyari	1	Saladhuddin Amin	.918
6	Ramadan Shallah	1	Ezzit Raad	1	Jawad Akbar	.787
7	Ezzit Raad	1	Fadal Sayadi	1	Omar Khyam	.704
8	Abdul Koyair	1	Abdul Koyair	1	Waheed Mahmood	.407
9	Abdul Qassim	1	Anjem Choudary	.626	Abu Izzadeen	.390
10	Mohammed Salim	1	Abu Izzadeen	.562	Abu Hamza	.343

Other preliminary findings suggest meaningful differences in a network may emerge based on data *type*. For example, differences appear to exist between the content of public and private statements made by organization leaders. Preliminary analysis of private statements by AM leaders suggests the importance of the conflict in Kashmir that is not apparent in public statements. Findings such as this reinforces our notion that collecting additional primary source material from former members of groups like AM will produce valuable insights.

The results of these analyses are highly predicated on the quality of thesauri and data used. By constructing our thesauri to exclude general terms (e.g., ‘children’) and historic figures and events (e.g., the Prophet Mohammed), we improve our ability to focus on evolving authority structures and other characteristics of the network. As the project moves forward, we will compare network structures based on different data types – beyond the use of ethnographic data to verify computational analysis described here – with the ultimate goal of triangulating data types as a means of cross-validating a comprehensive model of competitive adaptation.

3 Next Steps

As our computational modeling team continues to analyze the data we have provided to them on Al-Muhajiroun, our field researchers are collecting primary source data from former members of AM as well as from other groups. One field researcher recently completed two months of field work in Britain during which he conducted 39 interviews with former AM and splinter group affiliates. Interviewing such individuals can be done safely, ethically and does produce valid, policy-relevant data [1]. Using the theoretical framework of competitive adaptation, our research team will analyze these data with the objective of understanding how many of the internal processes affecting militant organizations affect their behavior, their use of terrorism (or lack thereof) and how they learn and adapt within their complex-adaptive environments. The resulting model of competitive adaptation will contribute an evidence-base to inform decision-making and law enforcement training, in addition to evaluating and forecasting the impact of specific policy interventions.

References

1. Horgan, J.: Interviewing the Terrorists: Reflections on Fieldwork and Implications for Psychological Research. Political Psychology (forthcoming 2011)
2. Jackson, B.A., Baker, J.C., Chalk, P., Cragin, K., Parachini, J.V., Trujillo, H.R.: Aptitude for Destruction Volume 1: Organizational Learning in Terrorist Groups and its Implication for Combating Terrorism. RAND, Santa Monica (2005)
3. Hamm, M.: Crimes Committed by Terrorist Groups: Theory. Research and Prevention. U.S. Department of Justice, Washington (2005)
4. Hamm, M.: Terrorism as Crime: From Oklahoma City to Al-Qaeda and Beyond. New York University Press, New York (2007)
5. Kenney, M.: From Pablo to Osama: Trafficking and Terrorist Networks, Government Bureaucracies, and Competitive Adaptation. Penn State Press, University Park (2007)
6. Wiktorowicz, Q.: Radical Islam Rising: Muslim Extremism in the West. Rowman & Littlefield, Lanham (2005)
7. Horgan, J., Taylor, M.: The Provisional Irish Republican Army: Command and Functional Structure. Terrorism and Political Violence 9, 1–32 (1997)
8. Hanneman, R.A., Riddle, M.: Introduction to Social Network Methods. University of California Riverside, Riverside (2005)
9. Carley, K.M.: On the Evolution of Social and Organizational Networks. In: Andrews, S.B., Knoke, D. (eds.) Networks In and Around Organizations: Special Issue of Research in the Sociology of Organizations, vol. 16, pp. 3–30. JAI Press, Inc., Stamford, Greenwich, CN (1999)
10. Jackson, B.A., Baker, J.C., Chalk, P., Cragin, K., Parachini, J.V., Trujillo, H.R.: Aptitude for Destruction: Case Studies of Organizational Learning in Five Terrorist Groups, vol. 2. RAND, Santa Monica (2005)
11. Simcox, R., Stuart, H., Ahmed, H.: Islamist Terrorism: The British Connections. The Centre for Social Cohesion, London (2010)
12. Raymond, C.Z.: Al Muhajiroun and Islam4UK: The group behind the ban. Working paper. International Centre for the Study of Radicalisation and Political Violence (May 2010)
13. Kenney, M.: Organizational Learning and Islamist Militancy. National Institute of Justice, U.S. Department of Justice, Washington, DC (2009)
14. Kenney, M.: “Dumb” yet Deadly: Local Knowledge and Poor Tradecraft among Islamist Militants in Britain and Spain. Studies in Conflict and Terrorism 33, 1–22 (2010)
15. Carley, K.M., Diesner, J., Reminga, J., Tsvetovat, M.: Toward an Interoperable Dynamic Network Analysis Toolkit. Decision Support Systems 43, 1324–1337 (2007)

Preferred Temporal Characteristics of an Advance Notification System for Autonomous Powered Wheelchair

Takuma Ito and Minoru Kamata

The University of Tokyo, Japan
ito@sl.t.u-tokyo.ac.jp

Abstract. In a rapidly aged society, providing mobility aids such as motorized wheelchairs is becoming increasingly important. Such low-speed vehicles have recently been developed with autonomous locomotion capabilities. In order to enhance the security and safety offered by these vehicles, human-machine interfaces are needed to inform the rider about the path of locomotion that is being taken. In this research we developed a prototype steering interface that provides haptic information to the rider about the locomotion of the vehicle. Initial experiments using a powered wheelchair simulator were performed to study the most acceptable temporal characteristics of the system in terms of the timing of the information provided to the rider.

Keywords: Locomotion Advance Notification, HMI, Low-Speed Vehicle, Autonomous Locomotion, System Acceptance.

1 Introduction

Electric powered low-speed vehicles have become increasingly common because of the ageing population. For the elderly who have difficulties in walking, various types of mobility scooters have been developed. However, decreases in cognitive ability and judgment due to ageing make it impossible for the users of mobility aids to drive, even if they were able to drive previously. For the elderly with decreased abilities, low-speed vehicles with autonomous locomotion would be helpful.

To date, several robotic powered wheelchairs have been developed. Some powered wheelchairs have semi-autonomous locomotion functions to aid users, while others have fully autonomous locomotion functions without any user input. The semi-autonomous ICW [1] developed by Kentaro K. et al. coordinates with terminals embedded in the environment, and has collision alert and avoidance functions. Meanwhile, the robotic wheelchair developed by Yoshinori K. [2] et al. uses visual information taken from the rider's face as well as the surroundings in order to move autonomously. Similarly, the TAO Aicle [3] developed by Osamu M. et al. coordinates with an infrastructure-based sensor system to travel autonomously. These projects have developed techniques necessary for autonomous locomotion, such as localization and object detection. To realize a secure and safe low-speed vehicle that is fully autonomous, other techniques for providing riders with a feeling of safety are necessary.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 394–403, 2011.

Advance notification of operating status is an effective means for providing a feeling of safety. Backup alarms on heavy trucks are typical examples of such notification, although the heavy trucks are not autonomous machines. Existing research dealing with advance notification of autonomous machines includes the research by Ryoujin I. et al. [4], who studied an advance notification system for the motion of a robotic arm, and demonstrated that the notification system was effective in suppressing the threat of injury to humans. In a related work, Takafumi M. [5] developed four types of advance notification systems for mobile robot locomotion and discussed their applications. All these researchers have focused on techniques for providing the people around the machine with a feeling of safety.

The goal of this research is the development of an interface that provides a feeling of safety to the rider of a low-speed autonomous vehicle. Because the rider of an autonomous vehicle is one of the persons around the autonomous machine, advance notification of the vehicle's locomotion should also be effective for the rider, to provide him or her with a feeling of safety. However, because the effectiveness of an advance notification system is expected to depend on the method of information presentation, appropriate design of the device and methods for providing information are necessary. Although there are many factors that should be considered for such design, this research focuses on the temporal characteristics of the advance notification system. Therefore, the main purpose of this research is the development of a prototype advance notification device, and to validate it in the perspective of how early should the locomotion system be notified. Although the intended final users of this system are the elderly with decreased cognitive ability, evaluation by the elderly at the initial stage is difficult for the variety of their characteristics. Therefore, in this study, the system is evaluated by the healthy young people as the first step. The outcome of this research will be one component of a safe and secure autonomous low-speed vehicle. Moreover, knowledge about temporal characteristics will be helpful for the design of sensing algorithms and autonomous locomotion algorithms.

2 Advance Notification System

2.1 Conceptual Design

For designing the user interfaces providing information, it is necessary to consider various types of information that are useful when presented to the user. Regarding this aspect, D. A. Norman [6] discussed and proposed six design rules for intelligent machines. Among these rules, the second rule: 'Be predictable' seems most important. Because autonomous vehicles do not rely on any user input, riders cannot know how and when the vehicles move. This unawareness has two disadvantages; unawareness of a locomotion plan can make riders feel anxious as well as prevent them from preparing for oscillations due to acceleration, deceleration or turning. An advance notification system would alleviate these concerns. In addition, riders had better to understand the state of the vehicle in case operations of the emergency stop are needed for the breakdown of the vehicle.

Regarding the interface of an autonomous vehicle, Norman's fourth rule, 'Make the output understandable', is also an important consideration. For example, for the situation shown in Fig. 1 in which a pedestrian is approaching the rider, the reason why the vehicle needs to avoid the pedestrian can be easily understood by the rider. Because the reason for an autonomous avoidance manoeuvre is obvious in this situation, presenting information to the rider is not critical. However, for the situation shown in Fig. 2 in which a bicycle approaches from behind, the rider has no way of knowing why the vehicle executes an avoidance manoeuvre. As unexplained autonomous locomotion sometimes makes riders feel anxious, reasoning information is also necessary. Therefore, not only locomotion information but also information explaining the reasons for locomotion are important.

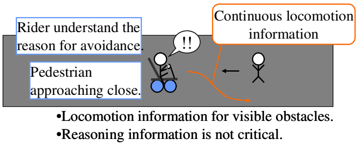

Rider understand the reason for avoidance.
Continuous locomotion information
!!
Pedestrian approaching close.
•Locomotion information for visible obstacles.
•Reasoning information is not critical.

Fig. 1. Acceptance of Locomotion Information for a Visible Obstacle (left side)

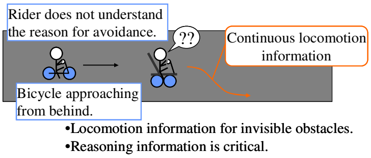

Rider does not understand the reason for avoidance.
??
Continuous locomotion information
Bicycle approaching from behind.
•Locomotion information for invisible obstacles.
•Reasoning information is critical.

Fig. 2. Acceptance of Locomotion Information for an Invisible Obstacle (right side)

Fig. 3 shows a schematic of the information system proposed in this research. First, on the basis of the information obtained from the sensors on the vehicle, the system alerts the rider that some type of action, such as an avoidance manoeuvre or deceleration, is necessary. Then the system announces the locomotion plan in advance before the vehicle actually moves. In this study, as the first step of this research, paths for turning are considered. In addition, to avoid any loss in meaning, it is necessary to separately provide both locomotion information and reasoning information. A method for separately informing the rider about the reasoning information has been discussed in a previous report [7]. Therefore, this research focuses on providing continuous locomotion information and explains the implementation of the system. The relationship between the users' acceptance of the information and the temporal characteristics of the locomotion advance notification system is also discussed.

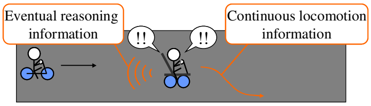

Eventual reasoning information
!!
!!
Continuous locomotion information

Fig. 3. Support Information Consisting of Reasoning Information and Locomotion Information

2.2 Implementation

There are several possible modes of communication for providing riders with various types of information. As each mode of communication has its own advantages and disadvantages, it is necessary to select an adequate mode of communication. Because our target autonomous wheelchair is expected to drive in pedestrian areas such as crowded sidewalks, the characteristics of the driving environment must be considered for designing the advance notification system. Because it would be difficult to use auditory devices and visual devices in this situation, a haptic device was selected. Specifically, as shown in Fig. 4, an Information Steering system consisting of a servo motor and steering grip was constructed as the prototype system. To inform the rider of the locomotion plan, information about the future yaw rate of the vehicle is assigned to the rotation of the steering grip. This assignment follows a metaphor for steering a conventional vehicle and therefore embodies Norman's third rule: 'Provide good conceptual models' and sixth rule: 'Exploit natural mappings'. Therefore, this prototype is intuitive and easily understandable. To examine the prototype's behaviour, the system was installed on a powered wheelchair driving simulator, as shown in Fig. 5.

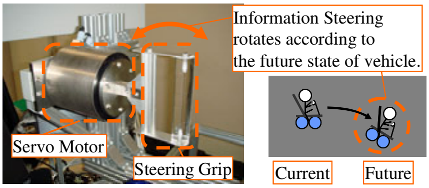


Fig. 4. Information Steering System

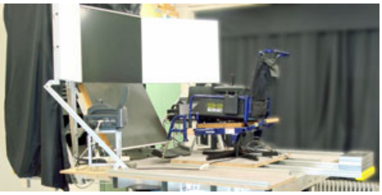


Fig. 5. Powered Wheelchair Driving Simulator

To make full use of this system, it is important to decide how early should the rider be provided with the information about locomotion. If the system provide the rider with the information about locomotion too late, the information would seem useless. On the other hand, if it is too early, the information would seem suspicious. Furthermore, considering the dominant factor in acceptable advance notification timing is necessary to develop a versatile device. As shown in Fig. 6, the main candidates for the dominant factors of acceptance are 'Distance' and 'Time', although there might be other candidates. In the existing research regarding risk indices for automobile collisions [8], the dominant factor seems to be 'Time'. However, the driving situations and the target vehicle in this research are different from those considered in the existing research for automobiles. Therefore, we conducted an experiment for examining the temporal characteristics of the advance notification system by studying riders' acceptance of the system on a driving simulator.

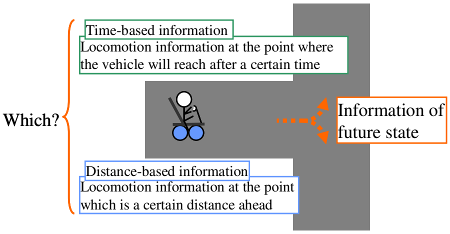

Time-based information
Locomotion information at the point where the vehicle will reach after a certain time
Which?
Information of future state
Distance-based information
Locomotion information at the point which is a certain distance ahead

Fig. 6. Preview Characteristics of Acceptable Information

3 Experiment

3.1 Overview

The purpose of this experiment was to evaluate subjective assessments of the temporal characteristics of the advance notification system. Subjects were made to experience autonomous locomotion on a powered wheelchair driving simulator and were supplied information about locomotion using the advance notification system equipped with an Information Steering system. The riders' acceptance of the information supplied was studied by changing the velocity, course layout and preview characteristics of the Information Steering system.

3.2 Method

Fig. 7 shows the layout and appearance of the experimental course. This course was short enough to conduct many trips without subjects' mental load, and the subjects turned only twice per trip in this course. The autonomous locomotion started from the point circled in orange, and turned at each corner. Because the direction of turning was randomly decided, the subjects did not know the direction of turning before being informed by the advance notification system. Table 1 shows the experimental conditions in terms of course length 'L', shown in Fig. 7, and velocity. These conditions were designed to be proportionate with the passing time so that the acceptable preview characteristics could be studied in terms of time and distance.

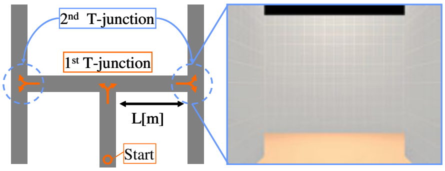


Fig. 7. Layout and Appearance of the Experimental Course

Table 1. Experimental Conditions

Condition	L [m]	Velocity [km/h]
I	9.0	4.5
II	4.0	2.0

In each trip, the preview time of advance notification of information about locomotion was changed between the two turns. The preview time at the second turn was 0.6 s shorter or longer than the preview time at the first turn. Fig. 8 shows an example of the changing preview time. Nine combinations of preview time were setup: from (0.0 s and 0.6 s) to (2.4 s and 3.0 s) stepped up by 0.3 s. The experiment consisted of four parts. Parts 1 and 3 were conducted with condition I, and parts 2 and 4 were conducted with condition II. Each part consisted of 18 trips: nine trips for which the preview times became shorter and nine trips for which the preview times became longer. The experimental order for testing each condition was randomized.

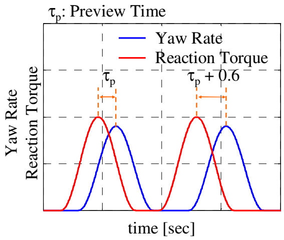


Fig. 8. Example of Changing the Preview Time

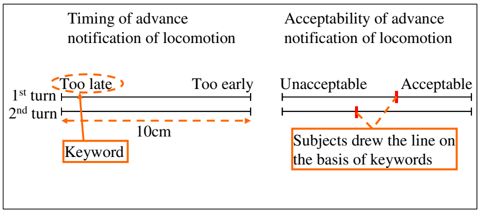

Timing of advance notification of locomotion
Acceptability of advance notification of locomotion
Too late
Too early
1st turn
2nd turn
10cm
Keyword
Unacceptable
Acceptable
Subjects drew the line on the basis of keywords

Fig. 9. Example of VAS Answer Sheet Used in the Experiment

After each trip, the subjects answered questions about the acceptance and timing of the advance notification of locomotion with a Visual Analogue Scale (VAS) [9]. Fig. 9 shows the VAS answer sheet used in this experiment. Two horizontal lines are drawn for each evaluation and the length of each line is 10 cm. Keywords such as 'Too early' and 'Acceptable' are written on both sides of each line. Instructions given to the subjects were as follows:

- Indicate your subjective evaluation of the acceptance and timing of the advance notification of locomotion by drawing a vertical line based on the keywords.
- Do not draw a line at exactly the same position for both turns.

After the completion of all trials, the subjects answered to a question sheet regarding various topics concerning the experiment. In this paper, the answers to following topics are discussed in the next section:

- What was the dominant factor of acceptance?
- How did you feel when the system presented information too early?
- How did you feel when the system presented information too late?
- How would your basis for evaluation change if the situation was more dynamic?
- How would your basis for evaluation change if the wheelchair was imperfect?

The subjects for this experiment were 14 university students who were publicly recruited to participate in the experiment. Although the intended final users of this system are the elderly with decreased cognitive ability, healthy young people were examined as the first step.

Before the experiments, the subjects experienced real driving with a real powered wheelchair. First, they drove a powered wheelchair by themselves. Then, they rode a powered wheelchair which was controlled remotely from a distance by the experimenter. This preliminary experience enabled the subjects to imagine riding an autonomous powered wheelchair. Subsequently, with the driving simulator, the subjects experienced some sample trips that included cases in which the preview time was the earliest and the latest. This sample trip was intended to prevent the subjects from changing their basis for evaluation during the main experiment.

3.3 Results

Although the answers on the question sheet indicated that most subjects evaluated the system positively, its effectiveness seemed to depend on the preview characteristics, as expected. Typical comments after the completion of all trials were as follows.

- 'If the system informed me too late, I felt I was in danger of colliding with the wall.'
- 'If the system informed me too early, I sometimes felt anxious.'
- 'If the system informed me too early, I could not judge whether the locomotion would be adequate or not.'

Because these comments indicated a strong dependence of the acceptance of the system on time, the distribution of acceptable preview times were analyzed numerically. The distributions for acceptance were approximately classified into three patterns:

Group A: Providing information too late and too early were both not acceptable. (eight subjects)
Group B: The earlier the information was provided, the better they felt. (three subjects)
Group C: No tendency. (three subjects)

Fig. 10 shows a typical distribution example for Group A. In the range of acceptance points, 0 indicates that the subjects drew the line on the left side and 100 indicates that they drew the line on the right side. Each value on the line chart is an average value of the evaluations for the conditions for which the preview time was same. A simple parabolic distribution can be observed and the maximum value exists within this range of data. The results for six subjects in Group A followed a similar distribution to varying degrees. On the other hand, the results for the other two subjects in Group A followed the distribution shown in Fig. 11; a parabolic distribution for condition I and a slope distribution for condition II. Because the upper limit for the preview time was 3.0 s in this experiment, based on the limited results, we could not conclude whether this slope distribution was the left part of the parabolic distribution. The answers on the question sheet after the completion of all trials were similar in the perspective that providing information too late or too early were both not acceptable. This result is consistent with the numerical parabolic distribution seen in Figs. 10 and 11.

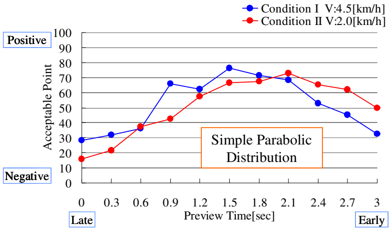


Fig. 10. Results for Subject S1 (Group A)

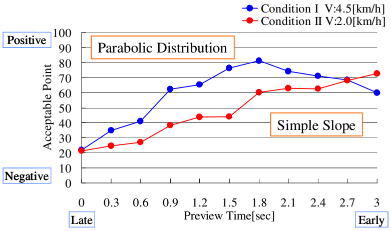


Fig. 11. Results for Subject S11 (Group A)

Fig. 12 shows a typical distribution for Group B. A simple slope distribution can be observed and the maximum value does not exist in this range. The answers on the question sheets from Group B were similar in the perspective that earlier information comforted the subjects. This result is consistent with the numerical parabolic distribution shown in Fig. 12.

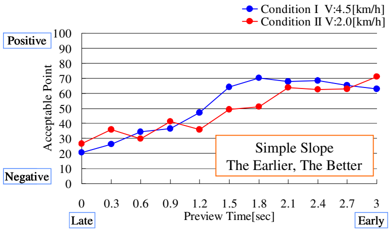


Fig. 12. Results for Subject S6 (Group B)

Figs. 13 and 14 show the distribution examples of the data for Group C. The results for this group did not show a clear trend. To comprehend these results, the VAS answers of three subjects, S8, S10 and S14, were analyzed in detail. Regarding the results from subject S8, the acceptance points for exactly the same condition were differed significantly between parts of the experiment. This indicates that this subject changed his basis for evaluation during the experiment. This change seems to be the cause of the absence of any trend in the data. On the other hand, the results for subjects S10 and S14 indicated that the correlation coefficient between the preview time and the subjective evaluation of the advance notification timing was very low. This indicates that they did not understand or answer the VAS sheet well. Inexperience with VAS seems to be the cause of the unclear distribution of the data.

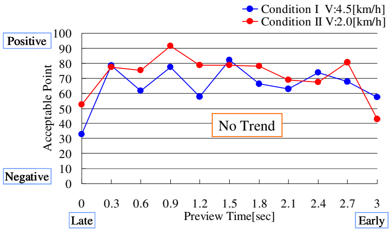


Fig. 13. Results for Subject S8 (Group C)

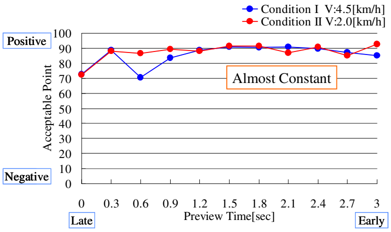


Fig. 14. Results for Subject S10 (Group C)

Based on the above results, it can be said that the most common trend in the results was parabolic. To understand this in more detail, the relationship between the trend in the data and the dominant factor for acceptance was analyzed. Table 2 presents a summary of the results for Group A. Because the acceptable preview times for both conditions are not inversely proportional to velocity, time seems to be the dominant factor. In fact, most subjects answered that time was one of the most dominant factors of acceptance.

The results for subjects S4, S11 and S13 seem particularly remarkable. Subject S4 answered that visibility around the corner was one of the dominant factors of acceptance, and subject S11 answered that distance, rather than time, was the dominant factor. In addition, the trend in the data for condition II was linear with a positive slope, and not parabolic. Furthermore, the difference of the most acceptable preview times between the conditions was greatest for subject S13, although the trend of this subject's responses to conditions I and II was parabolic. These results indicate the possibility that geometric factors could affect the evaluation basis of the acceptable preview time and cause the slope trend.

Table 2. Dominant Factors for the Acceptance of Locomotion Advance Notification and Most Acceptable Preview Time

		Dominant factor for the acceptance (Multiple answers are permitted.)			Most aceeptable preview time	
Subject	Trend	Time factor	Distance factor	Other factors	Condition I	Condition II
S1	Both parabolic	O	O		1.5	2.1
S3	Both parabolic	O	O		2.1	2.4
S4	Parabolic and **slope**	O		**Visibility**	1.8	X
S5	Both parabolic	O			1.8	1.2
S7	Both parabolic	O			0.9	1.5
S9	Both parabolic	O			1.2	0.9
S11	Parabolic and **slope**		**O**		1.8	X
S13	Both parabolic	O		**Visibility**	**1.2**	**2.1**

(Note: X indicates that local maximum value does not exist.)

In this experiment, it was assumed that the autonomous vehicle behaved ideally in a static situation without any interactions with pedestrians; these assumptions do not always hold true in the real world. To study the effects of these assumptions, the question sheet asked about the perceived effect of each assumption. The results indicate that all subjects answered that the acceptable preview timing should be less if the vehicle were moving autonomously in a crowded pedestrain area. In addition, most subjects answered that the acceptable preview timing would also need to be decreased if the autonomous vehicle were not perfect. These results indicate that the situation of the autonomous locomotion as well as degree of perfection of the autonomous vehicle are important factors.

4 Conclusion

This research aimed at developing a device for providing advance notification to improve the safety of autonomous low-speed vehicles. Based on the requirements of the system, a prototype Information Steering system was constructed. This prototype was examined with respect to riders' acceptance by using a powered wheelchair driving simulator. Experimental results indicated that advance notification seemed to be effective for riders of an autonomous low-speed vehicle. However, if the notifications came too early or too late, riders tended to feel anxious. Although the preview time seemed to have been the dominant factor for the acceptance of the advance notification system, geometric factors partially affected the acceptance as well. A preview time between 1.0 s and 2.0 s was acceptable for many riders.

However, we acknowledge that the driving situation used in this research was simple and static. In future study, experiments involving a more complex and dynamic driving situation are necessary. In addition, because the evaluations by the young subjects were conducted with using a driving simulator, implementation and evaluations by the elderly with a real vehicle also remain as a part of future study.

Acknowledgments. This work was supported by Grant-in-Aid for JSPS Fellows 10J08623. The author Takuma ITO was supported through the Global COE Program, “Global Center of Excellence for Mechanical Systems Innovation,” by the Ministry of Education, Culture, Sports, Science and Technology.

References

1. Kentaro, K., Ikuko, E.Y., Seiji, I.: Outdoor Environment Recognition System on Robotic Communication Terminals Supporting Mobility of Elderly and Disabled People. Systems and Computers in Japan 37, 56–67 (2006)
2. Yoshinori, K., Nobutaka, S., Yoshiaki, S.: A Robotic Wheelchair Based on the Integration of Human and Environmental Observations. IEEE Robotics & Automation Magazine 10, 26–34 (2003)
3. Osamu, M., Kiyoshi, K., Tsutomu, H., Tadao, Y., Shigeki, G.: Intelligent Wheelchair Robot TAO Aicle. In: Service Robot Applications, pp. 71–94. InTech, Rijeka (2008)
4. Ryojun, I., Yohei, K., Masatoki, I., Kazuki, M.: Previous notice method of robotic arm motion for suppressing threat to human. In: 2000 IEEE International Workshop on Robot and Human Interactive Communication, pp. 276–280. IEEE Press, New York (2000)
5. Takafumi, M.: Development of Four Kinds of Mobile Robot with Preliminary-Announcement and Indication Function of Upcoming Operation. J. Robotics and Mechatronics 19, 148–159 (2007)
6. Donald, A.N.: The Design of Future Things. Basic Books, New York (2007)
7. Takuma, I., Minoru, K.: Provision of Haptic/Tactile Information by a Vehicle Steering Interface. In: 19th IEEE International Symposium on Robot and Human Interactive Communication, pp. 27–32. IEEE Press, New York (2010)
8. Sou, K., Yoshitaka, M., Toshihiro, H., Makoto, I.: Comparison of Evaluation Indices concerning Estimation of Driver’s Risk Perception -Risk perception of rear-end collision to a preceding vehicle-. Review of Automotive Engineering 30, 191–198 (2009)
9. Maxwell, C.: Sensitivity and accuracy of the visual analogue scale: a psycho-physical classroom experiment. Br. J. clin. Pharmac. 6, 15–24 (1978)

Pre-validation of Nuclear Power Plant Control Room Design

Jari Laarni, Paula Savioja, Hannu Karvonen, and Leena Norros

VTT Technical Research Centre of Finland
Vuorimiehentie 3, Espoo, P.O. Box 1000, FI-02044 VTT, Finland
{jari.laarni,paula.savioja,hannu.karvonen,leena.norros}@vtt.fi

Abstract. Evaluation of the design of complex automation and control room systems is an essential phase in the design process in the nuclear field. For example, in order to meet the nuclear regulatory requirements, the new control room systems have to be evaluated in full-scope simulators to achieve a validation of the systems. We have developed a specific approach for the pre-validation of human-system interfaces and applied the method to evaluate the control room designs of a Finnish nuclear power plant. Some lessons learned from previous tests are provided. The paper will also discuss some open questions concerning the use of pre-validation test data. One of the most interesting questions is how pre-validation test data can be used in the final validation of a system, and how a set of pre-validation tests can support the validation by providing cumulative evidence of the functionality and usability of the system.

Keywords: Verification & Validation, Pre-validation, Control Room, Concept of Operations.

1 Background

Verification and validation (V&V) of complex automation and control room (CR) systems is an essential phase in the design process in many industries. For example, in the nuclear field, in order to meet the regulatory requirements, the new control room systems have to be evaluated in full-scope simulators to achieve a validation of the integrated system.

We have developed a specific approach for the integrated system validation (ISV) called Contextual Assessment of Systems Usability (CASU) [1]. ISV tests provide a final evaluation of the integrated system; before them, prototypes of the individual subsystems are evaluated through small-scale usability tests. Since complex systems are typically designed in a modular fashion, the designed systems are also evaluated in a stepwise manner. We have coined the term 'pre-validation test' to refer to these small-scale usability tests that precede the final testing. Even though there is a lot of literature on ISV testing, there is little research in the nuclear field on these pre-validation activities.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 404–413, 2011.

In this paper, we first describe the development of the pre-validation test methodology; then we give examples of the application of our approach and of how evidence of the system usability of the CR human-system interfaces (HSIs) is gathered based on the results of these tests. Typically, this phase also includes verification of design documents against standards and guidelines and verification of the design against design documents, but in this paper our focus is on the description of the pre-validation activities.

2 Development of Pre-validation Methodology

2.1 Requirements for Testing

There are some critical requirements for pre-validation testing. A key requirement is that the tests really support the iterative design of a system. Therefore, they should be carried out cost-effectively and quickly enough so that the input is delivered to the design process without delay. But at the same time they should be such that they truly assess the validity of the system i.e. they somehow remain independent from the immediate design solutions and also evaluate what kind of impacts a fully developed system would inflict in the future work. In addition, the tests should also be comprehensive enough to cover all the subsystems and their functionalities.

Timing has to be carefully planned, and testing of the system should be scheduled at the right time, i.e. at the moment when the design work has not yet been completed, and there is still time to take into account the recommended changes and fix them. If simulator tests at an engineering and design (E&D) simulator are included in the assortment of pre-validation activities, the plant model implemented in the simulator must also have been developed to a sufficient stage.

2.2 Selection of Pre-validation Methods

Different methods and techniques can be used in the evaluation of systems usability. Some of the most typical techniques are usability test, expert evaluation, cognitive walkthrough, focus group and usability questionnaires. We have adapted and refined most of these techniques to the evaluation of systems usability of complex technical systems, and they form the basis of our method assortment. The primary aim of a usability test is to improve the usability of a technical system through practical tests with users [2]. The participants who use the system represent real users, and they do real tasks. The personnel who are accomplishing the test observe and record what the users do with the system, how they communicate and how they co-operate. The data is analyzed and possible problems are identified. Based on the analysis, recommendations are given on how to improve the system and fix the problems. Expert evaluation is a kind of usability inspection method in which experts evaluate a HSI with a reference to a specific set of criteria, identify and rank the usability flaws according to their severity. In cognitive walkthroughs possible end-users of the system go through a sequence of actions with the user interface that is tested and evaluate its functionality and usability. Focus groups are group discussions in which participants tell their experiences and opinions about the usage of the system. Focus groups are quite often used in the early phases of the design work to probe possible users' attitudes and beliefs. A usability questionnaire is a list of items asking questions of the usability of a system.

2.3 Measuring Systems Usability

By applying the above-mentioned methods and techniques different aspects of human performance can be measured. It is assumed that if a system is usable humans can perform well with it. Typical factors of human performance evaluation are, e.g., personnel task performance, situation awareness, workload and teamwork. We have recently presented a slightly different classification of measures, based on the idea that different measures provide evidence of different perspectives of behavior. Performance measures are quantitative measures that give information of the outcome of human activity; practice measures give information of the core-task orientedness of activity; and user experiences give information of, e.g., the promisingness of the system for future work. Each of these classes of measures is needed when estimating the functionality and usability of a complex technical system.

3 Pre-validation Test Description

Five main phases can be identified in the pre-validation of CR HSIs, planning, modeling, data collection, analysis and assessment (Fig. 1). In the following, these five phases are described from the perspective of the consultant that is responsible of the evaluation of a new design.

3.1 Planning

In the planning phase the aim is to get familiar with the system that will be evaluated, formulate goals and constraints for the evaluation, and define the relevant methods and measures that will be used.

Training and Familiarization. Training of participants and personnel conducting the test is carried out before testing. Designers provide training on the new concept of operations, on new features of the HSI and on modifications of operational procedures. Technical feasibility of the simulator runs is also tested beforehand. It is important that demonstrations and simulations that are used in training are different from those that are used in actual simulator tests.

Defining Goals and Concerns. An important task in the planning phase is to determine the main focus of the pre-validation activities: 1) what systems are included in testing; and 2) from what perspective they are studied. For example, since operating procedures are typically tested together with the new HSIs, the target of a pre-validation test is not only a specific HSI element, but a set of operational activities that can emerge in a particular task.

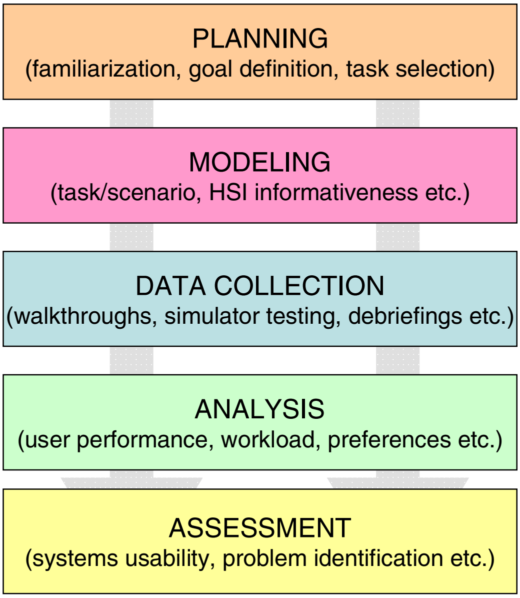

PLANNING
(familiarization, goal definition, task selection)
MODELING
(task/scenario, HSI informativeness etc.)
DATA COLLECTION
(walkthroughs, simulator testing, debriefings etc.)
ANALYSIS
(user performance, workload, preferences etc.)
ASSESSMENT
(systems usability, problem identification etc.)

Fig. 1. Main activities of a pre-validation test

Task Selection. The main task in pre-validation testing is the functional testing of HSIs in a simulator environment. Tasks and scenarios should be selected from the point of view of the systems to be tested. They should cover all the features of the HSIs that are included in the test, and a representative set of situations should be selected. If possible, there should be small-scale tests of particular features of the HSIs, and large-scale tests of the whole system. Ideally, the task selection is carried out in collaboration with designers, process experts and usability experts. But quite often, in the pre-validation phase, the usability experts do not participate in the selection of tasks.

3.2 Modeling

Since simulator testing of the proposed systems is a central task in the pre-validation of CR HSIs, tasks and scenarios for the simulator tests are modeled. In the modeling, the aim is to develop a conceptual basis for the assessment, and understand, analyze and describe the task-specific requirements for operator activity. A test scenario or a task is hierarchically analyzed to specify the task structure. For example, in a hierarchical task analysis, the main goal of the task is divided into the sub-goals to achieve it, and they can be further divided into lower-level goals if needed. Operating procedures can be used in the development of the hierarchical task breakdown structure. After the breakdown of the tasks, it is defined what information is presented on different display screens and other HSIs at different phases of task execution.

3.3 Data Collection

Recruitment of Personnel. The validation team typically consists of two-three human factors specialists of the consultant who conduct the tests and gather the data. Designers will participate in the training of other participants; they will also answer questions and provide additional information during the pre-validation activities. In addition, at least one simulator expert is needed to run the simulator. Typically, two or three crews of CR operators are recruited for pre-validation testing. If possible, it is preferable if operators with different levels of operator experience are selected.

Equipment and Material. Simulator models are developed in the E&D simulator before pre-validation testing, and pilot tests are conducted to verify the functioning of the E&D simulator.

All the material for briefings, walkthroughs, simulator tests and debriefings is prepared based on the modeling work. Detailed scenario descriptions are prepared for each test, and detailed guides for researchers are prepared before pilot testing. The guides include, e.g., instructions for the placement of test personnel and video cameras in the simulator, actions that are carried out, measures that are used and questions that are asked.

Description of Test Activities. The following research activities to collect information are typically used in a pre-validation test. The activities are presented in the same order they are usually conducted.

Observation of training sessions. Previous validation activities have shown that during training important usability issues emerge. Therefore, observation of training activities is an essential part of the validation process. The representatives of the validation team participate in the operator training sessions and gather comments that can be further discussed in the interviews.

Expert evaluation. Usability experts systematically evaluate the design before simulator testing focusing on general usability issues such as the visual layout of a user interface, and navigation properties and functionality of control devices.

Structured interview before simulator testing. All the operators who will participate in the simulator tests are interviewed beforehand. A special emphasis in the interview is placed on the evaluation of their knowledge and understanding of the new HSIs and/or concept of operations.

HSI-oriented walkthroughs. In order to evaluate the usability of the new HSIs, walkthroughs are carried out by using screen/paper mock-ups. Operators are asked about their positive and negative experiences with the new displays, and suggestions for improvements are gathered. They evaluate the design from the CR operator's point of view, concentrating on issues such as the possible lack of critical process information, and problems in the functional division of the system into display pages. A special emphasis in the walkthroughs can be put on those displays that play a small role in the simulator tests.

Simulator testing. In order to evaluate the new HSIs and/or the concept of operations, simulator tests are carried out. Simulator testing includes small-scale simulation tests with CR operators, in which, first, individual functionalities are tested, and after that, representative realistic simulator test runs are carried out.

The detailed structure of the test varies from test to test, but the basic structure of each test is as follows. Instructions to the operators are given including a short description of the status of the power process and automation system at the beginning of the test. Since the aim is to test the operators' ability to understand the new design and make operations with the new HSIs, it is important that the instructors do not provide answers to those operational tasks that are tested, but the operators themselves have a possibility to try to find the solution to the questions and do the operations that are needed. During the implementation of the test, members of the validation team make observations, video-record the test and rate online the performance.

Process tracing interview. A process tracing interview is carried out immediately after the test: The performed test is enacted and discussed with operators, and a set of questions is asked on the usability of the new design. Overall, the aim is to clarify the perception of the state of the process on which the operator's actions are based. In the interview, operators are asked to describe, e.g., 1) the process events that occurred in the test run; 2) what operations were associated with particular events; 3) what the meaning of each event was from a holistic process point of view; and 4) what information or user interface element the detection of a particular event was based on. The questions are modified to suit each specific task.

Questionnaires. After the complete simulator runs, the operators also complete the workload questionnaire, and after all the simulator tests, they also complete a usability questionnaire providing information of the functionality and usability of the new systems. The questionnaire includes statements about the usability of the CR. The statements include items of the control room's 1) instrumental function (e.g., task effectiveness), 2) psychological function (e.g., efficiency and suitability for the user), and 3) communicative function (e.g., support for shared situation awareness and cooperation). The participant evaluates how well the statement holds true by checking one of the four options.

Debriefing interview. At the end of each test day, a debriefing interview is arranged with operators, designers and usability experts. A special emphasis in the interview is placed on the evaluation of the role of the new operating system in the operator's work.

3.4 Data Analysis

In the analysis phase the data is processed in several successive phases. Pre-validation test data is analyzed mainly through qualitative analysis methods, but also quantitative analyses are carried out. Specifically, in the quantitative analysis of video data, behavioral research software is used. The video analysis is focussed on communications of the operators, directions of gazes, and operations and movements.

3.5 Assessment

The pre-validation activities provide both evidence of the validity of the concept of operations, the usability and functionality of a particular set of user interface elements and the adequacy of the training activities. In the following, these three areas are discussed separately.

Evaluation of Operational Concept. Both observational data and data gathered through interviews provide information of the effects of HSI changes on operator practices. Aspects of operator performance that are registered are 1) task completion (whether the operator could perform the action/task), 2) errors in performance (fault actions), 3) fluency of performance (amount and type of repetitions, interruptions and hesitations), and 4) communication and collaboration (number and content of speech acts).

By interviewing operators, we gather information of 1) operators' understanding of the concept of operations (e.g., their understanding of the function and meaning of the new systems), 2) operators' understanding of the differences between the new and old solutions, and 3) operators' situation awareness (concerning the status of the power process/automation system). Interviews also provide information of subjective experiences and preferences, e.g., subjective estimates of performance, situation awareness, and mental workload. They also provide information of the adequacy and promisingness of the new concept of operations, and recommendations and suggestions for improvements.

Based on the above-mentioned qualitative and quantitative evidence, an early assessment of the effects of the new HSIs on operator work practices is derived.

Evaluation of the Usability of HSI Components. Observations, interviews and walkthroughs provide evidence of the functionality and usability of the HSI components included in the scope of the pre-validation. HSI-oriented walkthroughs provide information of the main dimensions of usability, e.g., visual clarity, visibility, consistency, familiarity, flexibility and error prevention.

By observing operator performance we gather information, e.g., of task completion accuracy, fault actions and fluency of performance providing indirect evidence of the usability of the new design. Interviews, in turn, provide evidence of 1) operators' understanding of the use of information presentation formats (e.g., meaning of symbols and colour codes, logic of element groupings and display hierarchy), 2) user satisfaction with the new information presentation formats (e.g., use of symbols and colour coding, relevance and adequacy of element groupings and display hierarchy and navigation) in comparison to the old design, and 3) suggestions for improvements. Also the usability questionnaire that is completed at the end of the pre-validation session provides information of the functionality and usability of the new HSIs.

Based on the above-mentioned evidence, a preliminary assessment of the usability of the new design can be derived, and a list of possible problems and challenges can be prepared with suggestions for their solution.

Evaluation of Operator Training. Since operator interviews also give information of the relevance, adequacy and desired volume of training, some suggestions for the operator training can be given and a preliminary training concept can be outlined.

4 Example of the Application of the Method

A pre-validation test of a large screen display (LSD) pilot of a Finnish NPP was carried out at the E&D simulator in which the aim was to gather preliminary information of the usability of the prototype and gather experiences from participating operators [3].

Three members of the VTT research team carried out an expert evaluation of the displays before the functional tests. Since the expert evaluation was based on screenshots and paper images of the displays, the evaluation concentrated on basic design features of the displays. More complicated issues such as how these displays are integrated with the other displays in the CR and how they are controlled and managed were not considered.

Three crews of operators (pairs of operators) participated in the simulator test. For all the crews, the same set of six scenarios was provided. Detailed instructions for briefing the participants were developed. Before the usability test, a one-day training session was arranged, in which the aim was to familiarise the operators participating in the test with the key principles of the design concept and with the pilot, and to gather some first comments on the design solutions. During the first half of the training session designers gave presentations introducing the background and central ideas of the new display concept.

During the simulator runs the crews operated the plant as they were instructed and as they would have done in any other simulated or real situation. The whole scenario was recorded on four video tapes. Two of the video recordings provided an overview, and two of them were recordings produced with the head-mounted cameras. The evaluation team also took notes during the scenario in order to provide topics for discussion in later interviews.

In debriefing the main phases of the simulation were discussed through together with the operators. The aim of the interview was to find out what events the users considered most important in the simulated scenario, and what kind of information they used in order to manage the event. After all simulation runs, the operators were interviewed on their experiences about the displays that they have used in the test.

Some dimensions of operator performance (e.g., duration of time to event detection for each scenario, source of the first deviation detected and percentage gazing time to different information sources and number) were measured providing quantitative information of the use of LSDs. All the results of the pre-validation test activities were presented in a final report some weeks after the test sessions.

5 Conclusions

We have applied the pre-validation approach to several validation tasks in Finnish NPPs. Some of the lessons learned from these cases are the following:

- Pre-validation tests serve further development of the designed system. They provide information of whether the design work is proceeding according to agreed plans. Therefore, it has to be considered whether optimal design phase is selected for testing. The designed system must be complete and detailed enough in order for the testing to be feasible. In addition, the smooth functioning of the simulator model is important. It is hopeful that a large part of the target system has been simulated.
- Typically, in the pre-validation phase, systems are tested in a modular fashion, individual tests focus on a specific set of CR HSIs, and the HSIs are not tested as a part of the whole CR system. Since there is no certainty of the usability of the integrated system, we must be cautious in making inferences from pre-validation tests about the new concept of operations.
- The question of reference is a key issue in the evaluation of technical systems, since it is difficult to present arguments about the usability and functionality of the design, if there is nothing on which to base one's judgments. In the evaluation of individual features of HSIs the evaluation is based on the usability experts' judgment. In addition to that, standards and guidelines can be used. In the evaluation of concept of operations, the expertise of simulator trainers and experienced operators are needed, which may be difficult to obtain.
- Usability experts' independence from the design team is important also in the pre-validation phase, and its importance increases as the design process progresses. It is preferable that the usability experts are responsible for all the main activities of testing. In our cases, however, the representatives of the design team have been mainly responsible for the planning phase. Even though this does not compromise the reliability of the pre-validation, it is preferable if the usability experts could participate in the selection and modeling of tasks and scenarios.
- It seems to be that modeling and assessment are the most important and challenging phases in pre-validation. If there are not detailed enough models of the tasks, it is not possible to attend to key activities during simulation runs and ask relevant questions during process tracing interviews. In the assessment phase, a real challenge is to assess the safety implications of the design, or its impact on the concept of operations.
- In the assessment phase, our aim is not only to count usability problems and categorize them according to their scope and severity, but to evaluate how well the new system fulfils the functional criteria of systems usability (instrumental, psychological, communicative) [1]. In analyzing the instrumental function it is investigated to what degree the new systems support operational demands. In analyzing the psychological function it is evaluated how well the operators' coordination with the tools and procedures, and orienting to the core task demands, have succeeded. In the communicative function the focus is to judge whether the overall significance of singular events were comprehended and shared within the crew.
- A request that is included in our wish list is that a representative set of test scenarios are selected, and they cover all the tasks to which the new HSIs will be used. It is also desirable that a sufficient number of complete crews of operators with different levels of expertise are recruited, and the selected operators are a representative sample of the operating crews of the plant.

– Validation of NPP CR systems is typically considered as one distinct and integrated activity at the final stage of the design and Human Factors Engineering (HFE) process [4]. On the other hand, pre-validation activities are distributed along the design process, and they are tightly connected to many other activities (e.g., training, procedure design) of the HFE process. Therefore, we propose that a new kind of approach to V&V is needed, in which the evaluation is seen as a longitudinal and distributed activity in an integrated design process. The series of pre-validation tests conducted can thus support the more integrated validation of the CR HSIs by providing cumulative evidence of their systems usability.

References

1. Savioja, P., Norros, L.: Systems Usability - Promoting Core-task Oriented Work Practices. In: Law, E., Hvannberg, E.T., Cocton, G. (eds.) Maturing Usability: Quality in Software, Interaction and Value, pp. 123–143. Springer, London (2008)
2. Dumas, J.S., Redish, J.C.: A Practical Guide to Usability Testing. Intellect, Exeter (1999)
3. Laarni, J., Koskinen, H., Salo, L., Norros, L., Braseth, A., Nurmilaukas, V.: Evaluation of the Fortum IRD Pilot. In: Proceedings of the Sixth American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies NPIC&HMIT 2009. American Nuclear Society, LaGrange Park (2009)
4. O'Hara, J.M., Higgins, J.C., Persensky, J.J., Lewis, P.M., Bongarra, J.P.: Human Factors Engineering Program Review Model (NUREG-0711). NRC, Washington, DC (2004)

Deception and Self-awareness

Glyn Lawson[1], Alex Stedmon[1], Chloe Zhang[2], Dawn Eubanks[2], and Lara Frumkin[3]

[1] Human Factors Research Group, Faculty of Engineering, The University of Nottingham, Nottingham NG7 2RD, UK
[2] School of Management, University of Bath, Bath, BA2 7AY, UK
[3] School of Psychology, The University of East London, London E15 4LZ, UK
{glyn.lawson,alex.stedmon}@nottingham.ac.uk,
{kz222,d.eubanks}@bath.ac.uk, l.frumkin@uel.ac.uk

Abstract. This paper presents a study conducted for the Shades of Grey EPSRC research project (EP/H02302X/1), which aims to develop a suite of interventions for identifying terrorist activities. The study investigated the body movements demonstrated by participants while waiting to be interviewed, in one of two conditions: preparing to lie or preparing to tell the truth. The effect of self-awareness was also investigated, with half of the participants sitting in front of a full length mirror during the waiting period. The other half faced a blank wall. A significant interaction was found for the duration of hand/arm movements between the deception and self-awareness conditions (F=4.335, df=1;76, p<0.05). Without a mirror, participants expecting to lie spent less time moving their hands than those expecting to tell the truth; the opposite was seen in the presence of a mirror. This finding indicates a new research area worth further investigation.

Keywords: terrorism, deception, self-awareness.

1 Introduction

Recent statistics have shown that arrests associated with terrorism are rising, with 1,759 arrests occurring since September 11, 2001 [1]. In particular, intent to commit a terrorist act has increased by 30% since 2001. Terrorist attacks involving large-scale, high-value targets and widespread influences are considered strategic attacks, which involve a planning phase, including the processes of intelligence, surveillance and reconnaissance (ISR) [2, 3]. Most of the terrorist attack-planning indicators are hard to detect [3], and terrorists tend to behave differently based on their environments. However, at certain stages in ISR terrorists may be physically-present at their intended target [3] and need to conceal their intentions. This provides opportunities to identify suspicious individuals during the pre-attack stage using detection approaches.

The Shades of Grey research project aims to develop scientific interventions which will work on eliciting robust, reliable and operational indicators of suspicious behaviors, particularly relating to the reconnaissance stage of terrorist activities. This paper is associated with a work package which will develop and assess the value of different types of interventions, specifically aimed at revealing deception-related

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 414–423, 2011.

factors falling into the broad category of non-verbal behaviors. These behavioral cues might be aroused by intervention strategies designed to amplify suspicious reactions, in particular during reconnaissance of a terrorist attack in public areas.

Considering previous work on cues to deception, the Multi-Factor Model [4] proposes three factors which the influence behavioral cues to deception: emotion, cognitive effort, and attempted behavioral control. These factors also feature different aspects of deception, and the strength of such factors is highly relevant to cues associated with lying. These will be described below.

Ekman [5] argued that there are three different types of emotion associated with deception: fear, guilt and duping delight. Each factor that elicits emotional cues can occur all at once or in succession. Fear and excitement (the latter occurring through duping delight) might result in signs of arousal, such as increases in limb movements, speech fillers, and speech errors [6]. Guilt might result in gaze aversion [6]. Excitement may also result in signs of joy, such as smiling [7]. In spite of these deception cues, it is believed that liars try to use other facial expressions to mask signs of the emotion that they intended to conceal, in which case the effort of masking might fail [8]. Thus emotional leakage—which can be shown by facial expression [5, 9] or body movement—is a crucial non-verbal cue to deception.

Lying sometimes requires extra mental and cognitive effort than truth-telling. Because deceivers might be pre-occupied by formulating lies as well as remembering to play their role, they need to pay special attention to their behavior as well as monitoring the reaction of their targets, and they have to suppress the truth when they are lying. These processes for lying all require cognitive demand [6]. Deliberate efforts to "fight with" the conflict between lies and truth in their minds place mental demands upon liars e.g., [10, 11]. Evidence provided by neuroimaging studies (e.g., [12]) supports this point of view: the prefrontal cortex and anterior cingulate cortex are related to deception, which are involved in processing complex cognitive tasks and cognitive conflict. In addition, Carrión, Keenan, and Sebanz [13] revealed that tracing the target's mental state leads to greater cognitive demands compared with the conflict of the content of true or false statements.

People engaged in cognitive complexity present fewer hand and arm movements [8], less blinking [14], more [8] or less gaze aversion [15], and more speech hesitation [16] and errors [17]. They might display more pauses in speech, speak with a lower voice, and have longer reaction times, all of which are also found to be related to cognitive load [6]. The concentration which is aroused by cognitive overload thus influences behavior, such as the decrease in body movements, since the high cognitive demand leads to the neglect of body language [6]. As a consequence of cognitive overload, liars might be more rigid during deception (e.g., [17]; this is also caused by physiological inhibition from certain brain areas [18].

Concerning behavioral control, liars adjust their behaviors during lying by monitoring the reactions from their targets [19]. It is proposed that perceiving, monitoring and communicating with targets helps liars to successfully deceive (e.g., [20, 21]. Notably, in order to appear honest or normal, liars may attempt to control their behaviors during deception. Some evidence shows that liars may try to exhibit behaviors which they believe are credible, such as trying to behave positively and friendly to convince their targets [17]. However, such kind of deliberate self-regulation sometimes makes liars look over-controlled [6]. Some reviews [7, 17] indicate that liars' behavior might look rigid and tense, but speech might sound too smooth (presents less disturbances due to over control of speech) [16]. Furthermore, they might also be less forthcoming and less pleasant [17]. The complex presence of attempted behavioral control varies by person to person and it could be influenced by the simultaneous effect of emotion and cognitive load.

As described above, previous work has often been based on participants' behavior during interviews in which they are required to act deceptively. This study aimed to investigate cues to deception exhibited by people as they prepare to act deceptively in an interview. The outcome of this research could be used to support security personnel as they observe suspects prior to interview. Of relevance to behavioral control, this study also attempts to investigate the influence of self-awareness on deception cues.

2 Method

2.1 Participants

Recruitment was conducted by participant self-selection in response to posters and emails. Adverts specified that only undergraduate students should apply, and that they should not suffer from any mental ill-health. This requirement was to minimize the impact of any potential distress experienced from expecting to lie in the deception condition. 80 participants were recruited, 39 female and 41 male (mean age=20, SD=1.30, range=18-24).

2.2 Apparatus/Equipment

The experiment was conducted in a small office area. This contained a reception area and an interview area with chairs for the participant and the interviewer. All objects were removed from the walls to create an environment which was relatively free from distractions. There were no windows looking into or out of the interview area.

A camcorder was hidden within a green box file adjacent to the interview area. The aperture in the file had to be widened to enable the camera to capture then entire body of the participants. Because of this, it was possible to identify the lens, but only with close attention. Typical office products (glue stick, CD, marker pen) were located around the aperture to divert the participants' attention from the lens.

2.3 Experimental Design

The experiment took the form of a 2*2 between-subjects design. The two independent variables were:

- Self-awareness: mirror/no mirror. In the mirror condition, a full-length mirror was located directly opposite the participants in the waiting area. The mirror was removed and hidden for the no mirror condition.
- Deception: truth/lying. In the truth condition, participants were told to answer all questions truthfully. In the lying condition they were told that they could not answer any of the questions asked by the interviewer truthfully.

Thus, there were a total of four experimental conditions, of which participants were randomly assigned to one:

1. Mirror and truth (participants expect to answer truthfully)
2. Mirror and lying (participants have to invent answers and expect to answer untruthfully)
3. No mirror and truth
4. No mirror and lying.

2.4 Procedure

Participants were invited to take part in a trial to investigate deception skills in interview. Prior to each session, the hidden video camera was started. Upon arrival, participants were asked to sit in the reception area. A researcher explained to each participant that the study was being conducted to investigate deception skills in interview, and that after completing some preliminary forms and questionnaires an interviewer would arrive and ask questions about their degree courses. They were told that they should either answer truthfully (truth condition) or lie in all their answers (lying condition). Participants were asked to sign a consent form agreeing that they were willing to continue.

The researcher then led the participant to interview area. The researcher told the participant that they were leaving to find the interviewer and that they would return at the end of the session to complete participant payment forms. The researcher left the room under the pretence of going to find the interviewer. In reality, they hid outside the laboratory and timed five minutes. After this period, the researcher re-entered the room, apologized for the delay and asked the participant to return to the reception area. They told the participant that in fact there was not going to be an interview. After completing payment forms, and explaining the true purpose of the study, the hidden camera was stopped.

3 Results

The video footage of the 80 participants was coded using the Observer software. One researcher coded all footage. The coding scheme used is shown in Table 1. This was based on previous research into cues to deception, but was simplified due to the practical requirements for coding. Note that hand or arm movement includes any finger, hand or arm movement on either left, right or both sides; similarly foot or leg includes movement on left, right or both sides.

8 participants (10%) were randomly selected for coding by a second researcher to investigate inter-rater reliability. These were not used in the analysis of the behaviors, only to investigate the reliability. Cohen's Kappa, as calculated using the Observer software, was found to be significant, and towards the upper limits of "moderate" agreement (Kappa = 0.57; $p<0.01$). As the main results of interest included durations and frequencies these were also investigated. The durations were summed for the

movement categories for the eight participants. This was repeated for the second rater; the durations were found to be highly correlated between the raters (r_p=0.965, N=6, p<0.01). This process was repeated for the frequencies of the behaviors, which was also found to be highly correlated (r_p=0.923, N=6, p<0.01). Thus, the results were deemed sufficiently reliable for further analysis.

Table 1. Coding scheme

Hand or arm (either left or right)	Foot or leg (either left or right)	Whole body or torso	Gaze direction
Moving	Moving	Moving	Directly forwards
Still	Still	Still	Towards camera
			Other

The results are shown below, structured according to the movement categories in Table 1. Within each section the analyses are shown for the *duration* (i.e. total time spent moving) and *frequency* (i.e. total number of times the body part was moved regardless of duration) of movements.

3.1 Hand/Arm Movements

Hand and arm movement was first investigated using a 2*2 between-subjects ANOVA. The ANOVA for duration of the movements is shown in Table 2. This demonstrates a significant interaction between deception and self-awareness. The interaction plot is shown in Figure 1. This shows that without a mirror, participants expecting to tell the truth spend more time moving their hands than those expecting to lie; the opposite is seen in the presence of a mirror.

Table 2. ANOVA for duration of hand/arm movements

Effect	F	df	p	Eta^2
Deception level	0.034	1,76	NS	0.000
Self-awareness	0.280	1,76	NS	0.004
Deception*self-awareness	4.335	1,76	<0.05	0.054

The results of the 2*2 ANOVA for frequency of hand/arm movements is shown in Table 3. There were no significant main effects or interaction.

Table 3. ANOVA for frequency of hand/arm movements

Effect	F	df	p	Eta^2
Deception level	1.045	1,76	NS	0.014
Self-awareness	3.473	1,76	NS	0.044
Deception*self-awareness	1.305	1,76	NS	0.017

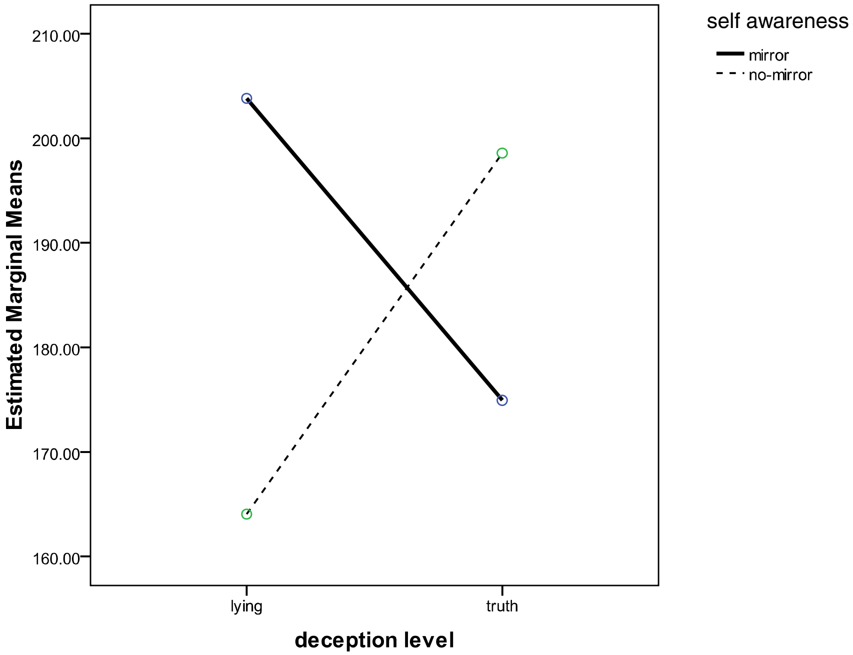


Fig. 1. Interaction plot for duration of hand/arm movements: deception*self-awareness

3.2 Leg/Foot Movements

No significant main effects or interactions were found for duration or frequency of leg/foot movements (Tables 4 and 5).

Table 4. ANOVA for duration of leg/foot movements

Effect	F	df	p	Eta^2
Deception level	0.386	1,76	NS	0.005
Self-awareness	0.807	1,76	NS	0.011
Deception*self-awareness	2.125	1,76	NS	0.027

Table 5. ANOVA for frequency of leg/foot movements

Effect	F	df	p	Eta^2
Deception level	0.767	1,76	NS	0.010
Self-awareness	0.971	1,76	NS	0.013
Deception*self-awareness	2.698	1,76	NS	0.034

3.3 Whole Body/Torso Movements

A main effect of self-awareness was found for duration of whole body/torso movements (Table 6). Those with the mirror spent longer moving (mean duration:

28.584s; SD=46.192) than those without the mirror (mean duration: 12.225s; SD=11.838). There were no significant findings for the frequency of whole body/torso movements (Table 7).

Table 6. ANOVA for duration of whole body/torso movements

Effect	F	df	p	Eta^2
Deception level	3.635	1,76	NS	0.046
Self-awareness	5.035	1,76	p<0.05	0.062
Interaction deception*self-awareness	3.789	1,76	NS	0.047

Table 7. ANOVA for frequency of whole body/torso movements

Effect	F	df	p	Eta^2
Deception level	0.137	1,76	NS	0.002
Self-awareness	3.423	1,76	NS	0.043
Interaction deception*self-awareness	1.232	1,76	NS	0.016

3.4 Gaze Direction

As gaze direction was a more complex measure than the previous behaviors, this was investigated using a 2*2*3 mixed ANOVA, with the variables of deception (expecting to lie/expecting to tell the truth), self-awareness (mirror/no mirror) and gaze direction (directly forwards/towards camera/other).

Table 8. ANOVA for duration of gaze direction

Effect	F[a]	df	p	Eta^2
Gaze direction	422.469	2,75	p<0.001	0.918
Direction*deception	1.577	2,75	NS	0.040
Direction*self-awareness	24.578	2,75	p<0.001	0.396
Direction*deception*self-awareness	0.251	2,75	NS	0.007
deception	0.616	1,76	NS	0.008
Self-awareness	4.698	1,76	p<0.05	0.058
Deception*self-awareness	0.614	1,76	NS	0.008

[a] Pillai’s Trace.

For durations, a main effect was seen for gaze direction, with most time spent looking at “other” (mean=205.681; SD=67.350) followed by “forward” (mean=66.914; SD=67.001) and finally looking towards the “camera” (mean=28.440; SD=23.957). The main effect for self-awareness was simply a result of measurement tolerances, and provides no meaningful data for understanding deception behavior.

A significant interaction for gaze direction and self-awareness can also be seen in Table 8. This finding indicates a change in gaze direction in the presence/absence of a mirror.

For frequencies, the main effect of gaze direction was found to be significant with the highest frequency for “other” (mean=11.92; SD=5.233) followed by “camera” (mean=7.37; SD=3.921) and finally “forward” (mean=6.79; SD=5.125). The interaction between direction and self-awareness was found to be significant, which also indicates a change in gaze direction in the presence/absence of a mirror.

Table 9. ANOVA for frequency of gaze direction

Effect	F^b	df	p	Eta^2
Gaze direction	114.462	2,75	p<0.001	0.753
Direction*deception	0.228	2,75	NS	0.006
Direction*self-awareness	12.595	2,75	p<0.001	0.251
Direction*deception*self-awareness	1.839	2,75	NS	0.047
Deception level	0.580	1,76	NS	0.008
Self-awareness	6.707	1,76	p<0.05	0.081
Interaction deception*self-awareness	1.412	1,76	NS	0.018

[b] Pillai’s Trace.

4 Discussion

This study indicated that few differences were observed in body movements between participants expecting to act deceptively and those expecting to tell the truth. The most notable finding was an interaction between self-awareness and deception for the duration of hand/arm movements: those in the lying condition moved their arms more in the presence of a mirror; the opposite was true for the truth tellers. It is difficult to understand why this interaction occurred, although it is certainly interesting that the presence of the mirror appears to magnify duration of the hand-arm movements of those expecting to lie. Previous research has demonstrated that people engaged in cognitive complexity (associated with lying) present fewer hand and arm movements [8].

The mirror resulted in an increase in whole body/torso movements. This may not be useful for identifying those expecting to lie, but contributes to an understanding of how people behave with increased levels of self-awareness. Similarly, gaze direction, and the interaction between gaze direction and self-awareness were significant, but these findings do not provide information with obvious use for detecting terrorist behavior.

Despite the finding that none of the body movements showed a main effect of deception level (truth telling vs. lying), the interaction in hand/arm movements suggests that the notion of self-awareness is worth further investigation as a possible tool for detecting deception. Future work could investigate in further detail the specific hand/arm movements in each condition (e.g. fold arms, tap fingers, touch face), to determine whether a certain type was more prevalent in each. For practicality this study used a high-level behavioral coding scheme, which could be broken down into further sub-categories for more detailed analysis. Behaviors could also be coded in a more subjective approach, for example focusing on behaviors associated with categories such as vanity, practice, nervousness etc.

Perhaps one further aspect to consider in future work is higher stakes. The only stakes in this experiment were participants' desire to convince the experimenter that they were telling the truth. With greater stakes the results may have been different [17].

5 Conclusions

This paper was an initial investigation into behaviors associated with deception while participants waited to be interviewed; previous research has generally focused on the behaviors demonstrated during an interview. This study also investigated the effects of self-awareness on cues to deception. An interaction was identified between deception and self-awareness for the duration of hand/arm movements (F=4.335, df=1;76, $p<0.05$). Liars moved their hands for longer when a mirror was present. This finding suggests that further research is required to understand the effects of self-awareness on non-verbal behaviors associated with deception, and in particular prior to the deceptive event itself. This research may ultimately improve the capability of security personnel to detect terrorists or people acting deceptively.

Acknowledgements. The authors acknowledge the EPSRC for funding this work.

References

1. Home office statistical bulletin. Retrieved from, `http://www.statistics.gov.uk`
2. Jessee, D.: Tactical means, strategic ends: al qaeda's use of denial and deception. Terrorism and Political Violence 18, 367–388 (2006)
3. O'Brien, K.A.: Assessing hostile reconnaissance and terrorist intelligence activities. The RUSI Journal 153, 34–39 (2008)
4. Zuckerman, M., DePaulo, B.M., Rosenthal, R.: Verbal and nonverbal communication of deception. In: Berkowitz, L. (ed.) Advances in Experimental Social Psychology, vol. 14, pp. 1–57. Academic Press, New York (1981)
5. Ekman, P.: Telling lies: Clues to deceit in the marketplace, politics, and marriage, 1st edn., pp. 43–161. Norton, New York (1985)
6. Vrij, A.: Detecting lies and deceit: pitfalls and opportunities, 2nd edn., pp. 1–188. Wiley, West Sussex (2008)
7. Memon, A., Vrij, A., Bull, R.: Psychology and law: truthfulness, accuracy and credibility, 2nd edn., pp. 1–55. Wiley, Chichester (2003)
8. Ekman, P.: Deception, lying and demeanor. In: Halpern, D.F., Voiskunskii, A. (eds.) States of mind: American and post-soviet perspectives on contemporary issues in psychology, pp. 93–105. Oxford University Press, Oxford (1997)
9. Ekman, P., O'Sullivan, M.: From flawed self-assessment to blatant whoppers: the utility of voluntary and involuntary behavior in detecting deception. Behavioral Sciences & the Law 24, 673–686 (2006)
10. Walczyk, J.J., Roper, K.S., Seemann, E., Humphrey, A.M.: Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. Applied Cognitive Psychology 17, 755–774 (2003)

11. Walczyk, J.J., Schwartz, J.P., Clifton, R., Adams, B., Wei, M., Zha, P.: Lying person-to-person about life events: A cognitive framework for lie detection. Personnel Psychology 59, 141–170 (2005)
12. Kozel, F., Johnson, K., Mu, Q., Grenesko, E., Laken, S., George, M.: Detecting deception using functional magnetic resonance imaging. Biological Psychiatry 58, 605–613 (2005)
13. Carrión, R.E., Keenan, J.P., Sebanz, N.: A truth that's told with bad intent: an ERP study of deception. Cognition 114, 105–110 (2010)
14. Bagley, J., Manelis, L.: Effect of awareness on an indicator of cognitive load. Perceptual and Motor Skills 49, 591–594 (1979)
15. Doherty-Sneddon, G., Phelps, F.G.: Gaze aversion: A response to cognitive or social difficulty? Memory & Cognition 33, 727–733 (2005)
16. Vrij, A., Heaven, S.: Vocal and verbal indicators of deception as a function of lie complexity. Psychology, Crime & Law 5, 203–215 (1999)
17. DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., Cooper, H.: Cues to deception. Psychological Bulletin 129, 74–118 (2003)
18. Vrij, A., Fisher, R., Mann, S., Leal, S.: Detecting deception by manipulating cognitive load. Trends in Cognitive Sciences 10, 141–142 (2006)
19. Buller, D., Burgoon, J.: Interpersonal deception theory. Communication Theory 6, 203–242 (1996)
20. Burgoon, J., Blair, J., Strom, R.: Cognitive biases and nonverbal cue availability in detecting deception. Human Communication Research 34, 572–599 (2008)
21. Burgoon, J., Buller, D., Floyd, K.: Does participation affect deception success? Human Communication Research 27, 503–534 (2001)

Air Passengers' Luggage Screening: What Is the Difference between Naïve People and Airport Screeners?

Xi Liu[1] and Alastair Gale[2]

[1] Civil Aviation University of China, Dongli Distriction, Tianjin, China, 300300
[2] Applied Vision Research Centre, Loughborough University, Loughborough, LE11 3UZ, UK

Abstract. In a simulated task of airport security inspection for threat items of knives, guns and IEDs, the difference between screeners and naïve people was analysed in terms of detection performance, attention allocation and workload. The detection performance of screeners was significantly better than that of naïve people. Compared to naïve observers, screeners concentrated on one or two potential threat items and ignored some irrelevant contents in the X-ray images which are showed by fixation maps. In order to understand how observers missed targets the workload between hit and miss decisions was compared. Unfortunately, there was no difference on workload when they hit or missed the targets where the dwell time on the targets of the hit decisions was longer than that of miss decisions. The findings may highlight how the search expertise is developed and provide information for improving training program.

Keywords: X-ray luggage image, visual search, fixation map.

1 Introduction

Currently airport screener performance degrades even further than the performance of 20% of potentially threat items missed in 1987 [1]. Other than the reasons such as large turnover, low wages and less experienced staff, monotony of the job and time pressure contribute this problem. Statistic data show that about 10^7 pieces of luggage are inspected every year at large international airports such as Heathrow [2]. This translates into an inspection time of about 6s per item which is consistent with the luggage screening time mentioned in a research paper [3]. The work environment of airport security screeners is noisy and images they search have a low signal-noise-ratio with varying background. For a success search screeners have to not only locate the proper area but also recognize the object even in a camouflaged situation or in a cluttered background. Furthermore, terrorists are inclined to make non-conventional explosives which are difficult to recognize. As a result of this, lots of advanced technology and equipment are used to assist screener to recognize threat items; but it is essential to understand the nature of the task and how the skill is developed and carried out since screener is still the decision-maker and executor of the search task.

The exact influence of experience and knowledge on search and recognition performance is not very clear. In radiographic interpretation domain, on one hand,

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 424–431, 2011.

studies [4] [5] showed that there are few significant changes in diagnostic performance after the first year of residency. Individual differences do not change with time and experience and performance is not consistent with the degree of experience and training. On the other hand, studies [6] [7] showed experts detect a target faster and search is completed more efficiently than less experienced observers. Novices' attention can be distracted by pseudo target and can be confused with potential mass and perturbation in breast parenchyma. In contrast, the non-experts used short interval, point by point examination of the chest radiograph with a more local scale, greater number of fixations and less scrutiny time [8]. Also, the visual search of screeners was more effective with a shorter time to first enter target areas and longer dwell time on the areas than these of naïve observers, especially for IEDs [9].

The study reported here exploits the difference of visual search and the allocation of visual attention between screeners and naïve people when they implement a simulated airport screening task for threat items of knives, guns and IEDs. Eye movement data analysis is employed to extract how observers allocate their attention on an X-ray luggage images which would highlight visual information processing and how screening expertise is developed. Measures of eye movement, fixations represent the location of conscious attention and saccades which describe temporal sequence of eye movement are too fast to gather information?. Fixations are relative to visual information acquisition and processing rather than saccades. A suspected target is often fixated subsequently or not sequentially re-fixated for verification in a search task and scan paths are observer specific [10]. Analysis of fixation duration, distribution and associated measures correlated with target areas in an image is a better opportunity to understand whether image features are selected than scan path although it gives fascinating insights into the temporal sequence of semantic interest [8]. In this study the eye movement data analyses primarily focus on the fixations for understanding how targets are hit or missed. Other analysis of eye movement data, such as saccade amplitude, blink frequency, blink duration and pupil diameter, is correlative to mental workload. Research shows that the increase in pupil size correlates with the increase in workload during visual search of symbolic displays [11], during performing an interactive route planning task [12], for the comparison of two weather displays by air traffic controllers [13]. Pupil size is combined with decisions of hit or miss to analyse the changes of workload level that could occur between different decisions in the present work. The aim of this study is to better understand the difference between screeners and naïve people in order to provide information for improving training program so as to enhance screeners' detection performance.

2 Method

2.1 Participants

Eight naïve people (2 female) and eight airport security screeners (4 female) participated in this study. All participants had normal or corrected-to-normal vision.

2.2 Stimuli and Apparatus

Thirty X-ray luggage pseudo colour images were collected for the study. Half of the images included a threat item: gun, knife or IED. The other half is normal. Images were displayed on a 21 inch monitor with a resolution of 1280 × 1024 pixels in 32-bit colour mode. Eye movements were recorded by a Tobii eye-tracker (X50) with temporal resolution of 50 Hz and spatial resolution of 0.35°.

2.3 Procedure

Participants were first calibrated for the eye tracker. They were familiarized with the task with several practice trials. The overall procedure was as follows. For each image, a visual noise mask containing a central fixation cross was first presented and participants were asked to fixate it. After 1 second it was then replaced by the stimulus image for unlimited time period. Participants interrupted the stimulus image presentation by pressing any key on the keyboard when they were confident of their decisions about whether a threat item is contained in the image. Then they were required to rate their decisions by a five point scale from 1 – 'definitely absent' to 5 – 'definitely present'. The location of potential threat item was indicated if the decision was 3, 4 or 5. After that, the participant can press any key for the next turn. All thirty images were presented in a random order.

Eight naïve participants were introduced to IEDs as being composed of a potential detonator, explosive, wires and a power source connected together before the experimental trials.

3 Results

3.1 Performance

Decision confidence data was analysed using the Receiver Operating Characteristic (ROC) method. The ROC curve, independence of the threshold chosen, is the graphic representation of this reciprocal relationship between sensitivity and specificity on the basis of all possible criterion value by sensitivity (true positive fraction, TPF) as y coordinate and false positive fraction (FPF = 1 – specificity) as x coordinate. Maximum likelihood estimation, the most widely used in medical imaging (Metz, 1986), was developed for estimating a smooth ROC curve based on binormal distribution. The area under such a fitted ROC curve, A_z value, is often used as an index to evaluate the diagnostic performance. In this study, A_z value was estimated using this method for representing detection performance of each participant in this simulated screening task. Figure 1 shows A_z values of screeners were higher those of naïve people. A t-test was carried out to compare the performance of screeners against naïve people and the result showed the performance of screeners was significantly better than that of naïve observers, $t = 3.77$, $df = 14$, $p < 0.01$.

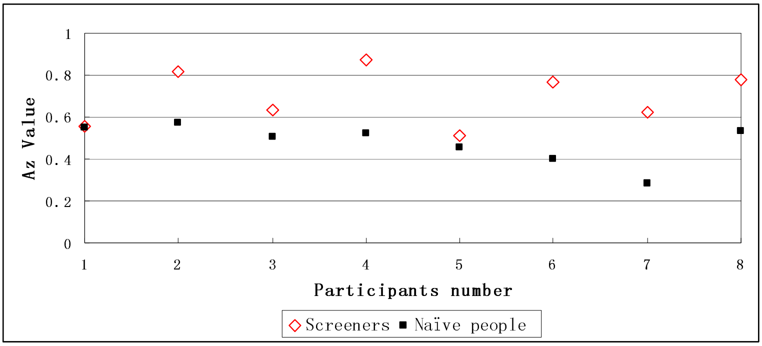


Fig. 1. The A_z value throughout 8 screeners and 8 naïve observers

3.2 Eye Movement Analysis

The variables of the time to first enter the AOI (area of interest) and dwell time on the AOI help people to learn about where people tend to look in an image and how people obtain visual information. The experience and expertise of screeners was showed by higher sensitivity with threat items which has been reported in another paper (Liu, Gale and Song, 2007). Other than the description measures, a method of fixation map was developed to intuitively express the distribution of fixations of one observer or a group of observers.

Pomplun, Ritter and Velichkovsky (1996) named the distribution map as attentional landscape, which was improved by Wooding (2002). The hypothesis of this method is that information falls off with distance from the centre of foveola, reflecting by the height of each fixation falls off with a distance from the centre of the fixation and a three dimensional Gaussian distribution centred at each fixation point. So each fixation is approximated by a Gaussian envelop. The height of each fixation is an indefinite value in Wooding's method while Pomplun and his colleagues (1996) weighed fixations for their durations. Wooding (2002) suggested the true width of the Gaussian is determined by stimuli and task requirement.

In our study, fixation map gives a chance to learn how people with different experience search X-ray luggage images based on spatial and temporal distribution of fixations. A MATLAB program has been developed to generate fixation map when the height and width of Gaussian envelop are determined. Here, the height of each fixation was weighed by its duration and the overlaps between fixations were accumulated into the final height at that point. The width of Gaussian envelop was decided by useful field of view (UFOV) of 2.5 degrees which was the filter of grouping raw eye movement data into fixation, since it represents characteristic of information processing of this search task (Gale et al., 2000).

The eye movement data of eight participants were pooled together and weighed the fixation duration to extract the distribution of all fixation data for each image. The height of final fixation map was determined by the duration of each fixation point and all existing fixations superposition at that map pixel. The height of certain fixation

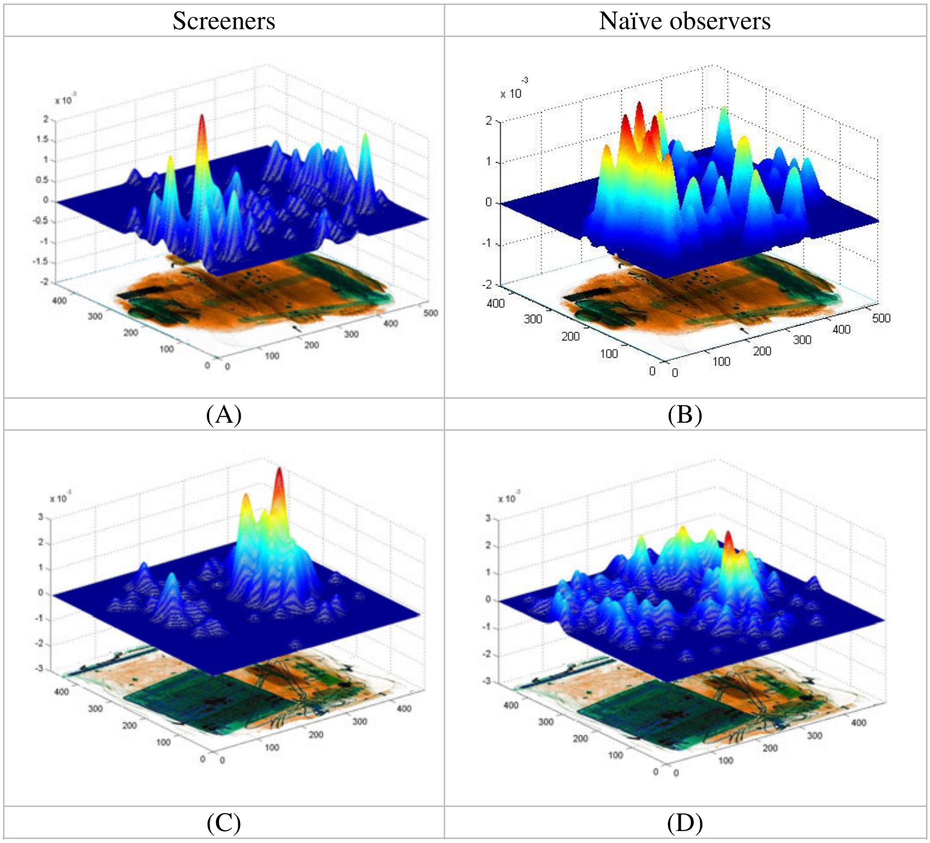


Fig. 2. Fixation maps of tow X-ray luggage images across eight screeners or eight naive people. A threat item was detected by six screeners (A), five naive people (B), eight screeners (C) and one naive people (D).

point was not the real fixation duration but it was in proportion to other fixation duration. Figure 2 is the example of fixation maps of two X-ray luggage images across eight screeners (A and C) or eight naïve observers (B and D). Fixation maps showed that observers were inclined to allot their attention on dark or dense areas, electronic components and wires.

Fixation distribution showed by Figure 2 (A) is a fixation map of a threat item in an X-ray luggage image detected by six screeners. Fixation distribution in Figure 2 (B) is a fixation map of a threat item in an X-ray luggage image detected by five naïve observers. For the luggage image, a gun from the end-on viewpoint is in the upper-left corner and several bottles are close to it. Screeners and naïve people fixated on them for a long time and the fixation height in the fixation map was the highest. Also they were inclined to pay their attention on the right of the image in which electronic components and wires are presented from the top down. The attention allocation tendency of screeners was on the left corner and the right corner of the image with longer dwell time while naïve people evenly alloted their attention all over the image.

It showed naïve people select more items to scrutinize and made a decision whether they were potential threat items when they searched targets in X-ray luggage images. Screeners were able to optimize their attention of ignoring some areas with no fixations or short dwell time on them. It seems screeners know where and what of search. This tendency was more obvious in the second image of figure 2.

Fixation distribution in Figure 2 (C) was a fixation map of a threat item in an X-ray luggage image detected by eight screeners. Fixation distribution showed by Figure 2 (D) was a fixation map of a threat item in an X-ray luggage image detected by one naïve observers. For the image, there is an electronic product on the left side, a pair of shoes on the right side and there are some hangers superposed on them. There are improvised explosive devices embedded in the shoes. Fixation map intuitively showed that screeners gazed at the target for a long time with an obvious peak. In contrast, naïve people scrutinized and compared more items so that the fixation map was composed by a lot of "hills" with similar height.

3.3 Mental Workload

Mean of pupil size of each observer was calculated for the comparison between screeners and naïve people. A t-test showed that pupil size of screeners was smaller than that of naïve people, $t = 4.356$, $df = 30$, $p < 0.001$. However, this could not demonstrate naïve people had a higher workload level than screeners when they implemented a search task for threat items since the pupil size baseline of observers did not measure before experiment task. If there was no difference for the pupil size baseline between two groups the result indicated that there was difference on workload between screeners and naïve people for the search task.

For screeners and naïve people, the dwell time on the AOI of hit decisions was longer than that of miss decisions, $t = 2.264$, $df = 14$, $p < 0.05$ and $t = 2.323$, $df = 14$, $p < 0.05$, separately. There was no difference for pupil size between decisions of hit and miss, $p > 0.1$, not only for screeners but for naïve people. Here pupil size was not sensitive enough to discriminate the difference of information processing and decision-making. There might be different in workload level for different decisions that could occur while other eye movement measures, including saccadic distance and blink duration, might capture it, which could be a future work.

4 Discussion

Participants of screeners and naïve people in the present study searched airport X-ray luggage images for threat items in a simulated airport security inspection task. The difference between two groups was analysed in terms of detection performance, attention allocation and workload. In contrast, screeners detected threat items in X-ray luggage images with higher hit rate and lower false alarm rate. The experience and ability of screeners were showed by better detection performance and more efficient eye movement search. Fixation map gives a chance to intuitively represent fixation distribution of two groups. Screeners allocated their attention to suspect areas of dark or dense objects, electronic products and wires. Also they concentrated on one or two locations for scrutinizing while naïve observers scanned all over the images and picked up several interest areas for comparing and recognizing. However, several targets still were missed even if they were fixated on. Familiarity with image and task led screeners to adopt a relatively simple search strategy which followed their experience and cognitive control. They were able to ignore some irrelevant contents in the images so that search was accelerated. Scanning patterns of naïve people showed by fixation maps indicated that attention was distracted by what the observer considered to be target candidates.

The difference between screeners and naïve people were discussed in terms of workload where pupil size was employed as a reliable measure. Unfortunately, the comparison in this study was ineffective since the baseline of pupil size before experiment trials was not recorded. For screeners and naïve people, there was no difference between the pupil size of hit decisions and the pupil size of miss decisions while dwell time on targets of hit decisions was longer than that of miss decisions. Extent of information processing was not reflected by pupil size. If there was difference on workload between hit and miss decisions, it might be reflected by subjective ratings or other eye movement activity measures. This would be our future work.

In summary, Fixation map analysis developed in this study is an objective method which quantifies parameters of eye movement, such as areas of interest and the degree of coverage. The eye movement difference of groups can be obtained by subtracting two fixation maps from each other. The difference of attention allocation patterns between screeners and naïve people showed by fixation maps provides a chance to learn how the expertise is developed although the factors that its development are not well understood. It would be a helpful attempt that attention of observers is guided to areas of potential threat items, which would increase the hit rate and improve the search efficiency. As a result, we suggest that eye movement trace and workload can be considered when a training program is designed or training interface is improved.

References

1. Singh, S., Singh, M.: Explosives Detection Systems for Aviation Security. Signal Processing 83, 31–55 (2003)
2. Speller, R.: Radiation-Based Security. Radiation Physics and Chemistry 61, 293–300 (2001)
3. Gale, A.G., Mugglestone, M., Purdy, K.J., McClumpha, A.: Is Airport Baggage Inspection just Another Medical Image? In: Krupinski, E.A. (ed.) Medical Imaging 2000: Image Perception and Performance. Proceedings of SPIE, vol. 3981, pp. 184–192 (2000)
4. Herman, P.G., Hessel, S.J.: Accuracy and its Relationship to Experience in the Interpretation of Chest Radiographs. Investigative Radiology 10, 62–67 (1975)
5. Gay, S.B., Hillman, B.J., McNulty, B.C., Altmaier, E.M., Smith, W.L.: The Effect of Pre-radiology Clinical Training on the Performance of Radiology Residents. Investigative Radiology 28, 1090–1094 (1993)
6. Krupinski, E.A.: Influence of Experience on Scanning Strategies in Mammography. In: Kundel, H.L. (ed.) Medical Imaging 1996: Imaging Perception. Proceeding of SPIE, vol. 2712, pp. 89–94 (1996)
7. Nodine, C.F., Kundel, H.L., Lauver, S.C., Toto, L.C.: Nature of Expertise in Searching Mammograms for Breast Masses. Academic Radiology 3(12), 1000–1006 (1996)

8. Manning, D.J., Ethell, S., Donovan, T., Crawford, T.J.: How do Radiologists do it? The Influence of Experience and Training on Searching for Chest Nodules. Radiography 12, 134–142 (2006)
9. Liu, X., Gale, A.G., Song, T.: Detection of Terrorist Threats in Air Passenger Luggage: Expertise Development. In: 41st Annual IEEE International Carnahan Conference on Security Technology, pp. 301–306. IEEE Press, Piscataway (2007)
10. Groner, R., Menz, C.: The Effect of Stimulus Characteristics, Task Requirements and Individual Differences on Scanning Patterns. In: Groner, R., McConkie, G.W., Menz, C. (eds.) Eye movements and human information processing. Proceedings of the XXIII International Congress of Psychology, pp. 239–241. North Holland, Amsterdam (1985)
11. Backs, R.W., Walrath, L.C.: Eye Movement and Pupillary Response Indices of Mental Workload During Visual Search of Symbolic Displays. Applied Ergonomics. 23, 243–254 (1992)
12. Iqbal, S.T., Adamczyk, P.D., Zheng, X.S., Bailey, B.P.: Towards an Index of Opportunity: Understanding Changes in Mental Workload during Task Execution. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, pp. 311–320. ACM Press, New York (2005)
13. Ahlstrom, U., Friedman-Berg, F.J.: Using Eye Movement Activity as a Correlate of Cognitive Workload. International Journal of Industrial Ergonomics. 36, 623–636 (2006)

Acceptability and Effects of Tools to Assist with Controller Managed Spacing in the Terminal Area

Lynne Martin[1], Michael Kupfer[1], Everett Palmer[2], Joey Mercer[1],
Todd Callantine[1], and Thomas Prevôt[2]

[1] San Jose State University
[2] NASA ARC: NASA Ames Research Center, Moffett Field, California, USA
{Lynne.H.Martin,Michael.Kupfer,Everett.Palmer,Joey.Mercer,
Todd.Callantine,Thomas.Prevot}@nasa.gov

Abstract. In a human-in-the-loop simulation, a scheduler delivered aircraft to meter fixes in the Los Angeles terminal area with a -60 to +30 second accuracy. This study investigated whether, and how well, controllers could control aircraft to land them as close to their scheduled time of arrival (STA) as possible using speed control alone. Controllers were assigned one of three levels of tools to assist them but had to compensate for errors in the forecast winds that had not been taken into account by the scheduler. Results show that speed clearances were sufficient under all conditions to maneuver aircraft closer to their STAs. From participant reports, this form of control incurred manageable workload and two of the three levels of tools were deemed easy to use.

Keywords: decision support tools, controller managed spacing, terminal area, utility and usability.

1 Introduction

One aim of the next generation air transportation system in the USA (NextGen)[1] is to maintain a high level of (or increase) the throughput at airports and more efficiently manage the traffic in dense terminal areas. Research in both Europe and the USA has focused on developing trajectory management tools that will enable aircraft to execute efficient descents and maintain throughput [e.g., 2, 3]. This requires more precise navigation, which is accomplished through more stringent Required Navigation Performance (RNP) criteria. Although NextGen Super Density Operations (SDO) [4] are founded in scheduling tools that can organize aircraft and specify scheduled times of arrival (STA), this arrival concept still requires controllers to work the traffic as it moves through the terminal radar approach control (TRACON) area to keep each aircraft on its tightly-packed schedule.

Currently TRACON controllers manually space aircraft around merge points and for landing, absorbing the frequent delays that are incurred from this control by speed changes and tactical vectoring (horizontal path changes)[5]. While vectoring is acceptable when aircraft are flying 3D paths, when operations become trajectory-based (TBO) – where aircraft fly 4D paths – vectoring is less efficient than speed

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 432–441, 2011.

control and contributes significantly to trajectory prediction uncertainty, along with wind variations. Realizing the benefits of the proposed scheduling tools requires a shift in current terminal area control practices away from vectoring strategies to strategic speed and path control on lateral area navigation (RNAV) routes (i.e., time-based procedures).

For TRACON controllers, this shift to TBO will mean not only a change in the way they control aircraft but also a shift in both the salient and critical information they will need to make control decisions. While at a casual glance, changing information when the task remains fundamentally the same would seem not to pose a problem, Nunes [6] cautions that some air traffic control (ATC) decision aiding tools have had adverse effects because they removed information. That controllers are very sensitive to key pieces of data is supported by Seamster, et al. [7], who found that expert air traffic controllers were able to efficiently focus on the most critical information for control decisions. Endsley and Rodgers [8] also found that ATC experts either pass over or do not commit less important data to memory.

Given the level to which ATC skills are honed and with a change in mode of aircraft control, it is likely that TRACON controllers will need to develop new strategies for their decision-making [7] and possibly revise aspects of their mental models [6] of the way traffic moves through the airspace. To support controllers with this, tools are proposed that assist at different levels of automation [9], from displaying schedule information to suggesting solutions (advisories). By introducing tools that provide different levels of information and assistance to ATCs, a comparison of tool effectiveness was facilitated.

The objective of the current study was to determine, through a human-in-the-loop simulation, how well controllers can manage spacing of arrival aircraft on Optimized Profile Descents (OPDs) along RNAV/ RNP routes with the assistance of different advisory tools and enhanced displays designed to assist them with keeping aircraft on their 4D trajectories. The trajectory-based tools were designed to support TRACON controllers by providing information for aircraft speed control that met a time-based schedule. Rather than testing each tool individually, the study investigated whether the tools could effectively be integrated into a terminal-area controller workstation, and how well the different tools functioned in concert.

2 Methods

2.1 The Simulation: Airspace, Route Structure, Scenarios and Winds

The airspace simulated for this study was the terminal area around the Los Angeles International Airport (LAX). Aircraft in the simulation flew OPDs on merging RNAV routes to runways LAX24R and LAX25L. The RNAV routes were designed based on existing Standard Terminal Arrival Routes (STARs; e.g., RIIVR2) and approaches. Several speed and altitude restrictions were created to give a sufficiently shallow descent angle of 2.4°. This allowed speed control to be used along the OPDs. The speed restrictions supplanted tactical controller speed assignments for fly-ability and predictable flow control. Fig. 1 shows a map of the simulated airspace displaying the routes, waypoints, and sector boundaries (based on current sectors in the Southern

California (SoCal) TRACON). The simulated airspace was comprised of three feeder sectors, Zuma, Feeder and Feeder South, and two final sectors, Stadium and Downe.

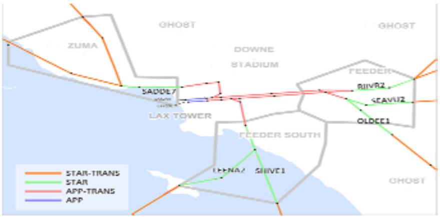


Fig. 1. The LAX airspace created for the simulation

Two different scenarios were developed for the simulation. Both scenarios included 25 aircraft flying to each runway. The aircraft type mix and the traffic load distributions on the routes were selected based on an analysis of actual arrival traffic to LAX. The scenarios were built under the assumption that aircraft had been delivered to the TRACON meter fixes by en route control with no more than a 90 second nominal schedule error (up to 60s early and 30s late). However, due to the wind forecast errors, the error between the estimated time of arrival (ETA) and STA differed from that range. In addition to the standard wake spacing distances an additional buffer of 0.5 NM (Nautical Miles) was added into the scheduler to protect the wake spacing and reduce the possibility of violations.

Winds were always a headwind aligned with the landing runway from 265°. Above 20,000ft and below 1,500ft the forecast wind profile matched the actual wind profile. However between these altitudes there were two wind-forecast-error conditions where the actual wind differed from the forecast wind. This has an effect on the accuracy of the higher-level tools (see below) because their calculations take the forecast winds into account. In the "minus-bias" wind condition the forecast winds were 10kts less than the actual winds and in the "plus-bias" wind condition the forecast winds were 10kts stronger than the actual winds. A third "no-bias" condition was also used, where the actual winds were the same as the forecast.

2.2 Tools

The study was run in the Airspace Operations Laboratory (AOL) at the NASA Ames Research Center using Multi Aircraft Control System (MACS) software [10]. MACS provides an environment for rapid prototyping, human-in-the-loop air traffic simulations, and evaluation of current and future air/ground operations. Simulated aircraft were assumed to be Flight Management System and Automatic Dependent Surveillance-Broadcast-out equipped.

The controllers worked with an emulation of the Standard Terminal Area Replacement System (STARS) onto which one of three levels of decision support tools (DST) was added. The first level toolset could be implemented in the near-term NextGen; it consisted of a double-sided timeline comparing aircraft STAs with their ETAs. The timeline (TL; Fig.2.1) was referenced to an appropriate waypoint for each controller, e.g., LAX25L for the Downe position. Accompanying the timeline was an early/ late indicator displayed in the third line of an aircraft's data block (FDB) if the aircraft was five seconds or more early or late. These DSTs showed a controller how close an aircraft was to being on time and therefore the amount of time that s/he needed to create or absorb to put the aircraft on schedule.

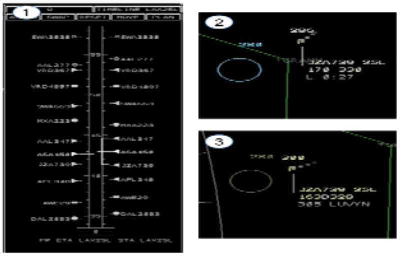


Fig. 2. In clockwise order from left: (1) timeline, (2) slot marker and early/late indicator in slot marker condition, (3) advisory and slot marker in advisory condition

The second level toolset added position prediction information in the form of a slot marker (SM) to the schedule information. The SM showed with a circle where the aircraft should be on its longitudinal arrival route in order to fly the charted speed profile in the forecast winds and arrive at the runway on its STA. This means that an aircraft in the center of its slot marker circle would be properly spaced behind its lead (providing the lead aircraft is also in the center of its slot marker). This level of support provided a target around the best schedule position for an aircraft that a controller could work towards.

The third level toolset may be envisaged in the far-term NextGen; this toolset retained the timelines and the slot marker circles but replaced the early/late indicators in the FDB with speed advisories (AD). The speed advisories suggested a speed over a distance, which, if the aircraft flew it, would put the aircraft onto its STA at the outer marker. Like the early/late indicator, the speed advisories were displayed to the controllers if an aircraft's ETA differed from its STA by five seconds or more. This DST is at a higher level of automation [9] than the other two conditions because the tool recommends a solution to the controller rather than leaving the controller to work out a solution on his or her own.

2.3 Controller Tasks, Participants, and Data Collection

The goal set for our participants was to efficiently deliver aircraft on their routes to the outer marker and runway with no wake spacing violations. The controllers were asked to manage the arrival traffic, correcting those aircraft ahead or behind schedule, and to cope with disturbances (i.e., aircraft not conforming to speed restrictions), all while dealing with the errors between forecast and actual winds. They were asked to do this using only speed control if possible. The role of the feeder controllers was to issue an approach clearance for each aircraft along an RNAV/RNP route to its assigned runway, then try to deliver the aircraft as close as possible to its STA by the sector exit point.

The final controller was tasked with further fine-tuning the traffic received from the feeder controllers to deliver the aircraft at their STAs at the outer marker.

Five air traffic controllers took part in the study. They had an average of 23.6 years of experience and had been retired for 1.9 years on average. Confederates who were also retired controllers staffed two ghost positions. The pseudo-pilots who worked the traffic were active commercial pilots or aviation students who had experience using the MACS software.

Prior to data collection, participants received three days of training in fully running practice simulations where they worked the position they would work during the data collection. Data was collected during simulation runs over five consecutive days. Each tool and wind condition combination was run under the two traffic scenarios to give a total of 18 runs. One run had to be repeated due to a procedural error. Each run lasted for an hour and immediately following it participants completed a questionnaire about that run. At the end of the data collection, participants completed another questionnaire about more general topics, and took part in a debrief discussion.

Each controller and pseudo-pilot workstation recorded a number of variables in data logs throughout every simulation run. Aircraft performance data, trajectory and flight state information as well as pilot and controller data entries were logged. As an extra task, controllers were prompted, at five-minute intervals, to give a rating of their workload for that moment between 1 ("very low") and 6 ("very high") through a workload assessment scale based on the ATWIT (Air Traffic Workload Input Technique [11]) that was embedded in the MACS software. Voice communications between controllers and pilots were through an emulation of the FAA's Voice Switching and Communication System and these communications were recorded.

3 Results

3.1 Task and Goal Achievement

A number of metrics were calculated to assess whether the controller team achieved their goals. Route conformance is one: in order to receive the benefits of OPDs aircraft are required to conform to their route with high precision. This was achieved; in all conditions route conformance within 1 NM was approximately 99.5%. Moreover, the controllers did not use vectoring techniques to manage the arrival stream, completing their tasks using speed control alone. Vertically there were very few level-offs, only 19.2% of aircraft leveled off for an average of 2.48NM. Most of these were due to deceleration for waypoint restrictions, which again means that controllers adhered to the OPDs as much as they could.

The second goal set for the controllers was to deliver aircraft with minimum spacing between a lead aircraft and its follower when crossing the runway threshold. Overall, the average inter-arrival spacing of aircraft pairs did not vary much when compared across tool conditions; they varied around 0.53 NM ($\sigma = 0.27$ NM), which reflects the spacing buffer. Throughout all of the data collection runs involving 900 aircraft and 864 aircraft pairs, there was only one wake spacing violation that was due to controller error. It took place during a timeline condition and the follower was

0.12NM closer to its leader than it should have been, suggesting controllers generally maintained a high level of traffic awareness.

A third goal was for controllers to control the aircraft in order to arrive at the runway on schedule, known as “schedule conformance”. This metric was defined as an aircraft’s ETA minus its STA (negative values indicate the aircraft was ahead of schedule). Schedule conformance indicates a combined effort from the controllers, because to achieve this goal the controller team must reduce the ETA-STA differences over the entire length of a flight through the TRACON. Compared to the initial –60 to +30 s distribution of traffic the schedule errors measured at the runway are greatly reduced. The distribution peaks around -5 s with a mean of μ=-1.21 s and standard deviation of σ = 5.21 s (Fig. 3). The curve is sharper on the left, indicating controller effort not to exceed the schedule buffer, and is wider to the right – excess spacing is somewhat inefficient but keeps aircraft separated.

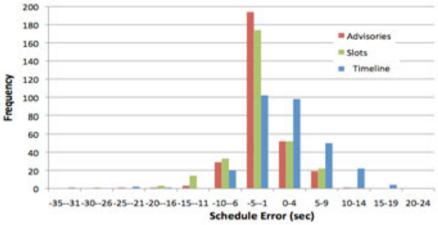


Fig. 3. ETA-STA error measured at the runway threshold broken down by tool condition

Breaking down the results by tools condition shows that the advisories and the slot markers have very similar schedule conformance, while the timeline condition has a less tight conformance to the schedule. A one-way ANOVA indicated that these conformance patterns are significantly different (F(2,838)= 48.32, p=.000). A post-hoc Tukey’s HSD test confirms that all the tool conditions are significantly different from each other at the p<.05 level. From this it seems the pattern of schedule conformance is most different in the timeline condition while the slot marker and advisory conditions’ patterns are more similar.

3.2 Task Load and Its Acceptability

In addition to the metrics of task achievement, another set of measures was taken to assess whether the load on the controllers was reasonable while they were completing the study tasks. Controller workload was measured in real-time using an ATWIT-based procedure [11]. All controllers perceived their workload on average as “low” to “very low” (μ = 1.85, σ = 0.53). Although the response scale offered six choices, controllers effectively used only a 3-point scale – from the raw scores, no controller ever gave a rating of 5 or 6 and rarely gave a rating of 4. Some of these workload ratings, those reported from the feeder controllers, stem from low sector traffic complexity because only arrival operations were simulated.

When participants’ workload ratings are distributed by the study’s conditions, there are only small differences between participants’ estimations of their workload (Fig. 4).

A comparison between the mean workload ratings of the three toolset conditions showed the mean rating for the timeline condition was higher than that for the speed advisory condition but only by .08 of a scale point (TL μ = 1.68, AD μ =1.60). Unsurprisingly, differences between participants' mean ATWIT ratings per run are not statistically significant when compared by tool condition.

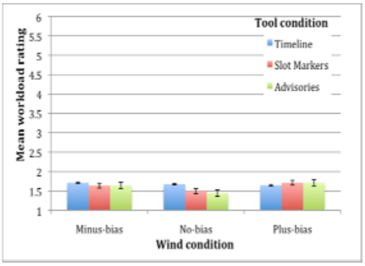


Fig. 4. Mean of ongoing workload ratings by study condition, with standard error bars

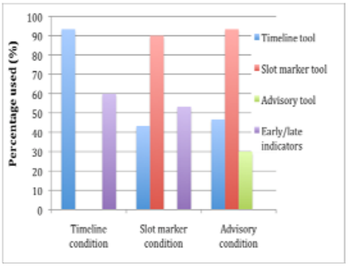


Fig. 5. Amount participants "used" the tools in each toolset condition

As a complement to the real-time workload ratings, workload data was also collected in post-run questionnaires using the NASA-TLX [12]. Controllers completed six scales that comprise this rating scheme after each run, using a ranking that ran from very low workload (1) to very high workload (7). Overall participants' average workload was between "low" and "somewhat low" (μ=2.57, σ=1.36). When the TLX scores were organized by tool condition the differences between the means were small, similar to the differences between the ATWIT means. Once again participants rated the timeline condition as having the most combined workload (μ = 2.64) and the slot marker and advisory conditions as having almost identical workload (SM μ=2.51; AD μ=2.56). Again, there were no significant differences between participants' ratings of their TLX workload. Considering that participants were only undertaking a portion of the tasks working these study sectors that they would normally have when working these sectors at SoCAL, workload ratings of "low' to "very low" are reasonable.

Controller load was assessed in a third way through the amount of communication that was required. All clearances were issued by voice, thus creating the physical load for participants and potentially a time constraint as voice clearances have to be issued serially. Across all runs aircraft received a mean of 2.5 clearances per controller. The controller of Zuma, exclusively handling aircraft flying to LAX24R, issued the most clearances on average (μ = 3.1, σ = 1.41) while, as expected, the Feeder South controller (LAX25L traffic only) issued on average the fewest clearances (μ = 1.98, σ = 1.24). At this broad level, the number of clearances issued support participants' physical load and time pressure reports; with the Zuma and Stadium controllers issuing the most clearances on average and reporting the highest mean physical load and time pressure, and the Downe and Feeder South controllers issuing the fewest clearances on average and also reporting the lowest mean workload. Comparing across the different tool conditions, the aircraft under the timeline condition received fewer clearances on average compared to the other two tool conditions (TL: $\mu = 2.18$, $\sigma = 1.4$; SM: $\mu = 2.70$, $\sigma = 1.42$; AD: $\mu = 2.64$, $\sigma = 1.37$). Comparisons of the mean clearances per aircraft do not have a statistical difference.

3.3 Tool Usage and Acceptability

Controllers were asked a number of questions about how the tools impacted the way they completed their tasks after each run. For example controllers were asked how much they used each of the four key tools and how useful they were. Note that the slot markers and the advisories were not available in every tool condition but the timelines were. Participants reported they used the timeline significantly more often ($\chi^2(2) = 8.897$, $p = 0.012$) in the timeline condition (93% of the time) when compared with the other two conditions where they reported using the timeline much less (46.6% and 43.3% of the time) (Fig. 5).

This would suggest that participants could use the timeline but it was not their first choice of tool. Although there was a statistical difference between the amount controllers said they used the slot markers in the slot marker condition versus the advisory condition, it was not a meaningful difference because controllers said they used the slot markers 93% of the time in the advisory condition and 90% of the time in the slot marker condition. However, this *is* meaningful in terms of the tools. Controllers reported they used the slot marker (as noted) 93% of the time in the advisory condition but they only used the advisories 30% of the time – which indicates they chose not to use the most advanced tool. Controllers' comments support that they preferred the slot markers over the advisories and the timelines as their "tool of choice". Controllers used the early/late indicators about the same amount in the timeline and slot marker conditions but more than they reported using the advisories that replaced them in the advisory condition.

In a post-study questionnaire, controllers were asked about how useful each tool was and how useable it was. Responses to these two questions are highly correlated ($\tau = 0.66$, $p < 0.01$) and show that, in general, if participants thought a tool was useful they also thought it had a high level of usability. The slot markers were rated as "very useful" ($\mu = 4.6$) with "high usability" ($\mu = 5$), and participants commented: "Using them helped [me] to make adjustments on aircraft" indicating that they were using the markers in the way the study intended them to be used. The same was true for the early/late indicators; participants used them as planned ("Used the early/late advisories until they would disappear then I would use the timeline for the final adjustment") and they were rated favorably as "useful" ($\mu = 4.2$) with "high usability" ($\mu = 5$). The timeline was also rated positively as both "useful" ($\mu = 4$) and "useable" ($\mu = 4.2$) and gets a third positive rating because it was the only one of the three main tools that no-one said they would have liked to have been able to turn off. However, participants did say that they found the timeline hard to use because it "took my attentions away from the radar screen." The speed advisory was rated lower and more variably than the other tools. Overall, participants said the advisories were "somewhat useful" ($\mu = 2.8$) and "somewhat usable" ($\mu = 2.75$) but, also, three participants would have liked to be able to turn them off or use them for information only.

Debrief discussions probed a little further into why the participants did not view the advisory tool as favorably as the other tools. Participants explained that the advisory tool often issued an advisory to a waypoint that was downstream of their sector, which meant (if the controller issued the advisory) that the aircraft would not be in conformance by the time it was due to be handed off to the next controller. This is at odds with current-day controller work techniques to complete all tasks related to a given aircraft before handing it off. Thus, controllers balked at issuing the speed advisory and instead tried to create a solution that would be completely implemented by the point at which they wanted to hand-off the aircraft.

4 Discussion and Conclusions

Participants completed their tasks and met the goals of the study relatively easily. The feeder controllers delivered a well-conditioned flow to the final sectors, and the final controllers merged aircraft and avoided excess spacing and wake spacing violations at the runways. They were assisted in their tasks by the support tools but the relatively low traffic load and lack of scenario complexity made the tasks easier to complete than anticipated. The lack of complexity was reflected directly in participants' low workload ratings and neither workload scale was sensitive enough to detect finer variations in controllers' workload reports that may have thrown light on whether some traffic was more difficult to manage than others and why.

Although the toolsets showed little relationship to controller performance (performance being high in all conditions), participants expressed a clear preference for the slot marker condition. They controlled the traffic by using the slot marker circles as a spatial control target with the early/late indicator, and then used the timeline for further fine-tuning. A comment controllers made more than once was that they would have liked the early/late indicator, which was displayed down to an ETA-STA error-precision of five seconds, to continue to be displayed until there was only one second of ETA-STA mismatch, i.e., until the aircraft was in its target position.

Despite being the most advanced tool offered in the study, the speed advisory was not perceived as more useful but actually as less so, and participants reported using it proportionately less than the other tools. This was because the advisories did not conform to controller norms and so controllers interpreted, rather than followed, them. Of course, interpreting advisories requires mental manipulation, which accounts for the similar levels of mental workload to other toolsets reported in this condition when it was expected to be lower. Because of this mismatch between controller work techniques, a concern that the more automated speed advisory tool may reduce controller understanding was not borne out. However, this is still potentially a hazard, as when the tool is fine-tuned to match controller strategies the potential for controllers to use the advisory without thinking may recur.

Lessons learned from this study that will carry forward into future controller managed spacing work are threefold. First, there is a need to refine the decision support toolsets, the most major being to consider allocating portions of the speed advisory on a sector-by-sector basis. These amendments have been developed and are presently being tested [13]. Second, consideration must be given to the need for controllers to maintain a deep level of understanding of the traffic that they are managing. Third, with respect to testing these more refined tools, it is possible that participants would lean on proffered tools more heavily if the scenarios were more complex or if larger disruptions to the schedule occurred.

In sum this simulation of merging terminal area arrival traffic showed that controllers, assisted by simple-to-use and informative decision support tools, were able to correct for initial schedule and wind forecasting errors and deliver aircraft on-schedule using just speed clearances. There were only small performance and workload differences when comparing between the DSTs. The preferred toolset was the slot marker toolset including timelines, slot marker circles and early/late indicators, which highlighted key information for the controllers but did not constrain their choice of strategies.

Acknowledgments. Thank you to the AOL development team and the controller subject matter experts without whom this study could not have been completed.

References

1. Joint Planning and Development Office: Concept of Operations for the Next Generation of Air Transportation System, Version 3.0. Washington, DC, October 1 (2009)
2. Coppenbarger, R., Dyer, G., Hayashi, M., Lanier, R., Stell, L., Sweet, D.: Design and test of automation for efficient arrivals in constrained airspace. In: Coppenbarger, R., Dyer, G., Hayashi, M., Lanier, R., Stell, L., Sweet, D. (eds.) International Congress of the Aeronautical Sciences, AIAA2006-774, Nice, France (September 2010)
3. Boursier, L., Favennec, B., Hoffman, E., Trzmiel, A., Vergne, F., Zeghal, K.: Integrating aircraft flows in the terminal area with no radar vectoring. In: Boursier, L., Favennec, B., Hoffman, E., Trzmiel, A., Vergne, F., Zeghal, K. (eds.) 6th AIAA Aviation Technology, Integration and Operations Conference (ATIO), Wichita, KS, September 25-27 (2006)
4. Isaacson, D.: Airspace Super Density Operations (ASDO) Concept of Operations. NASA Ames Research Center, CA (2007)
5. Shepley, J.: Near-term terminal area automation for arrival coordination. In: Eighth USA/Europe Air Traffic Management Research & Development Seminar, Napa, CA (2009)
6. Nunes, A.: The impact of automation use of the mental model: Findings from the air traffic control domain. In: Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society, Santa Monica, CA (2003)
7. Seamster, T., Redding, R., Cannon, J., Rider, J., Purcell, J.: Cognitive task analysis of expertise in air traffic control. International Journal of Aviation Psychology 3(4), 257–283 (1993)
8. Endsley, M., Rodgers, M.: Situation Awareness information requirements for en route air traffic control, DOT/FAA/AM-94/27. US DoT, FAA, Washington, DC (December 1994)
9. Sheridan, T.: Supervisory Control. In: Salvendy, G. (ed.) Handbook of Human Factors, pp. 1244–1268. John Wiley & Sons, New York (1987)
10. Prevôt, T.: Exploring the many perspectives of distributed air traffic management: The Multi-Aircraft Control System MACS. In: Chatty, S., Hansman, J., Boy, G. (eds.) HCI-Aero 2002, vol. 2002, pp. 149–154. AIAA Press, Menlo Park (2002)
11. Stein, E.S.: Air trafic controller workload: An examination of workload probe, DOT/FAA/CT-TN84124. Atlantic City International Aiport, NJ, FAA (1985)
12. Hart, S.G., Staveland, L.E.: Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In: Hancock, P., Meshkati, N. (eds.) Human mental workload, pp. 139–183. Elsevier, Amsterdam (1988)
13. Swenson, H., Thipphavong, J., Sadovsky, A., Chen, L., Sullivan, C., Martin, L.: Design and evaluation of the Terminal Area Precision Scheduling and Spacing System. In: Ninth USA/Europe Air Traffic Management Research & Development Seminar, Berlin, Germany (submitted)

The Effects of Individual and Context on Aggression in Repeated Social Interaction

Jolie M. Martin[1], Ion Juvina[2], Christian Lebiere[2], and Cleotilde Gonzalez[1]

[1] Dynamic Decision Making Laboratory, Department of Social and Decision Sciences
[2] Psychology Department,
Carnegie Mellon University, 4609 Winthrop Street, Pittsburgh, PA 15213, USA
{jolie,ijuvina,cl,coty}@cmu.edu

Abstract. In two studies using variations of the Prisoner's Dilemma game, we explore the combined impact of individual traits and social context on aggressive behavior. In the first study, we compared defection rates in the Iterated Prisoner's Dilemma when participants were presented with a payoff matrix (Description condition) or learned payoffs through experience (Experience condition). Interpersonal trust and maximizing tendency led to relatively more cooperation in the Description condition than in the Experience condition, demonstrating that individual characteristics manifest differently depending on the information available to decision-makers. In the second study, we employed a new game paradigm, the Intergroup Prisoner's Dilemma with Intragroup Power Dynamics, to examine the way that power motives influence extreme aggressive behavior. We discovered that certain individuals exhibit very high levels of defection, but only when they play with particular combinations of predefined strategies, suggesting further how the confluence of individual factors and context can induce aggression.

Keywords: Aggression, Extremism, Game Theory, Individual Differences, Power, Prisoner's Dilemma, Social Context.

1 Introduction

One of the primary difficulties of studying aggressive behavior is its rarity. While instances of aggression are common on the news, they are relatively sparse within the complete set of human interaction that takes place on a day-to-day basis. It is fortunate, of course, that most people behave in accordance with social norms of civility, and yet highly aggressive tendencies at the other end of the spectrum need to be understood for the widespread destruction they can cause even in rare occurrence. Such aggression is almost certainly rooted in some confluence of individual predisposition and perpetuation via social context [1].

Due to the relative infrequency of terrorist and other violent actions within the totality of actions by people toward one another, the preponderance of research in predicting extreme behavior typically examines either specific cases or large data sets from the real world, focusing on those instances where extreme behavior occurs. Here, we take a different approach by attempting to study aggressive behavior in a

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 442–451, 2011.

laboratory setting with "normal" populations where we would expect actual extremism to be unlikely. Rather, we look for tendencies toward aggression insofar as we can measure them in controlled environments.

Game Theory has been widely used to study behavior in conflict situations [2]. One of the benefits of this approach is that it distills complex real-world interactions into their core components that can be manipulated in a laboratory setting. We use such abstract games in this research to test individual and contextual drivers of aggressive behavior. In Sections 2 and 3, we describe two experiments using variations of the Prisoner's Dilemma game to illustrate the combined effects of individual background variables and social context on aggressive behavior. Then, in Section 4, we conclude with implications of this research for predicting extreme behavior in the real world.

2 Effects of Individual and Social Variables

The Prisoner's Dilemma (PD) game is one of the most common paradigms for studying conflict experimentally. Like other "2x2" games, it involves interaction between two players who make simultaneous choices between two available actions. In PD, these actions are often referred to as cooperate (C) and defect (D). Typical payoffs for each combination of player actions are shown in Table 1. The Nash equilibrium in PD is for both players to defect since this action gives each individual higher payoffs regardless of the other's action [3]. However, a unique tension arises in the PD from pursuing personal gain through defection versus increasing social welfare (i.e., joint payoffs) through mutual cooperation, and researchers have consistently noted well below 100% defection even in the simple one-shot PD [2].

Table 1. Typical PD payoff matrix with action C denoting cooperation and action D denoting defection. The cells show pairs of outcomes (payoff of Player 1, payoff of Player 2).

		Player 2 Action	
		C	D
Player 1 Action	C	1, 1	-10, 10
	D	10, -10	-1, -1

The Iterated Prisoner's Dilemma (IPD) played for multiple rounds between the same pair of opponents brings about additional tradeoffs. First, it may complicate a purely economic computation of personal payoff maximization, since defection may increase immediate payoffs at the expense of establishing trust and encouraging an opponent to cooperate, which generates higher payoffs later. Second, and perhaps more importantly, repeated interaction enhances social concerns such as fairness [4] and reputation [5]. Individuals may feel altruism or malice toward an opponent, reciprocity or retribution for an opponent's past actions, and the desire to signal trustworthiness or threat to induce future cooperation by an opponent. Since there may also be many equilibria in the infinitely repeated IPD (which, in practice, is finite but with no known endpoint for players), we focus on the way that these latter factors may influence aggression.

2.1 Experience versus Description in IPD

A primary objective of the present study was to explore the effects of varying social information on individual behavior in conflict. The methods and results are reported in greater detail by Martin et al. [6]. Here, we focus on understanding the combined impacts of an individual's background variables and social information in inducing aggressive behavior.

In one condition of the IPD, we gave participants descriptive information about their interdependence in the form of a payoff matrix; in another condition, participants only learned the outcomes of their joint actions through experience. Differences in decisions from description and experience have been widely documented in individual risky choice [7], yet this gap has not been extensively studied in adversarial interactions.

We hypothesized that social information would enhance players' inferences about the other's motivations, and – assuming higher-level reasoning takes place – what others might infer about their own motivations. In addition, we predicted that the information available would accentuate different aspects of individual traits, leading to interactions between individual and context in predicting aggressive behavior.

2.2 Method

We recruited 120 participants (age *M* = 24.9 years, *SD* = 7.07 years; 61% male) to the Dynamic Decision Making Laboratory at Carnegie Mellon University in multiple sessions. Participants were randomly matched in anonymous pairs to play 200 unnumbered rounds of IPD over the Internet. In each round, they simultaneously chose between actions labeled indistinctly as Action A and Action B, with payoffs in points as shown in Table 1. To create incentives for performance, participants were informed that their points would be converted to money at the end of the game.

To represent the disparities of information that decision-makers may have in real conflict, 30 pairs (60 participants) were randomly assigned to each of two conditions differing in the amount of information given to players. In the Description condition, as in most studies using the IPD and other 2x2 games, players were shown the complete payoff matrix from the outset of the game and throughout their interaction. They could thus refer back to this information about the structure of the game as a concrete reminder of the outcomes for each player as a result of both their actions. In the Experience condition, participants also saw the actions that the other player took in each round, as well as the other player's payoffs, but were not given the matrix. After several rounds were experienced, this condition would be objectively identical to the Description condition from a 'rational' standpoint: participants were aware what outcomes would accrue to each player based on their two actions in a given round. However, there was no matrix present to serve as a reminder of the game's strategic structure. We coded the extra information by condition (1=Description, 0=Experience) in order to explore interactions with individual variables.

Participants completed an online questionnaire to capture several background variables that have demonstrated relevance to decision making in social situations: a five-item interpersonal trust scale measured belief in the honesty and reliability of others [8]; a nine-item maximizing tendency scale measured strength of preference for attaining the best possible outcome over one that is merely satisfactory [9]. Both of these scales produce scores ranging from 1 to 5 based on a participant's average responses (from 1=strongly disagree to 5=strongly agree) on each question. We assess how each of these interacts with condition in predicting defection rate in the IPD.

2.3 Results

On average, participants scored 1.65 on interpersonal trust (*SD* = 0.71) and 3.48 on maximizing tendency (*SD* = 0.63). These are our independent measures of participant traits, with the experimental condition serving as an independent measure of information available in conflict.

The main dependent measure for aggressive behavior was the proportion of defection actions that a participant took throughout the 200 rounds. Figure 1 shows average defection rates in the two conditions. The distribution appears bimodal in both conditions, since many pairs reach a steady state of fairly consistent mutual defection or mutual cooperation, but there are a greater number of participants with high levels of defection in the Experience condition than in the Description condition. The average defection rate was 0.57 (*SD* = 0.35) in the Description condition, compared to .69 (*SD* = 0.29) in the Experience condition ($t(118) = 2.04$, $p = 0.04$).

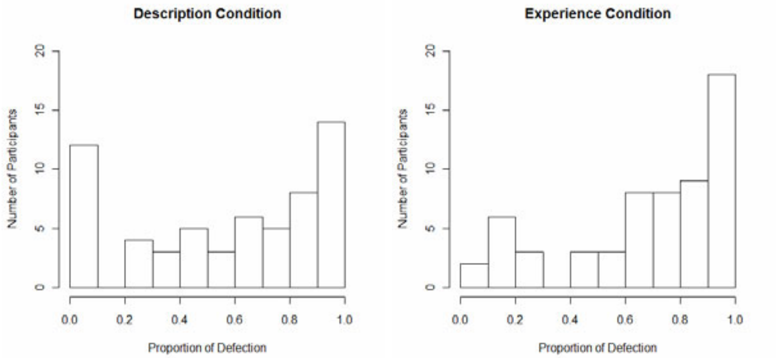


Fig. 1. Proportion of defection in the Description and Experience conditions

Of greater interest in the present analysis was whether individual traits played a different role in predicting defection across the two conditions. We fitted a regression model to predict a participant's average defection rate from interpersonal trust score, maximizing tendency score, experimental condition, as well as interactions between the individual variables and condition. Overall, these variables explained a reasonable amount of variance ($R^2 = .26$, $F(5, 114) = 7.95$, $p < 0.001$). When controlling for other variables, we observed a main effect of condition, with descriptive information leading to greater defection ($\beta = 1.21$, $t(114) = 3.72$, $p < 0.001$). There was also a marginally significant negative main effect of interpersonal trust (β = -.10, t(114) = -1.93, p = 0.06), and positive main effect of maximizing tendency (β = .15, t(114) = 2.84, p = 0.005). This suggests, not surprisingly, that those who were less trusting of others and those who sought to maximize their own payoffs defected more in the IPD. However, we also observed a marginally significant negative interaction between condition and interpersonal trust (β = -.14, t(114) = -1.86, p = 0.07), and a significant negative interaction between condition and maximization tendency (β = -.32, t(114) = -3.69, p < 0.001).

2.4 Discussion

We speculate that removal of the payoff matrix in the Experience condition increased defection more so for people who were inherently trusting because they became more emotionally reactive, hence taking defection personally and ascribing ill intentions to opponents, which is known as the attribution error [10]. At the same time, those participants with higher maximizing tendencies might have engaged in more localized processing of recent outcomes in the Experience condition, focusing on the immediate gains of defection and becoming especially loss averse with respect to the risk of a large loss from unilateral cooperation with a defecting opponent [11].

Conversely, it is possible that having the payoff matrix present in the Description condition encouraged more cooperation by increasing the ability of participants to develop a ‘theory of mind’ or impute mental states for one another [12]. For those who were inherently trusting of others, the matrix might have increased empathy for the symmetrical decision faced by an opponent, leading to greater forgiveness of defection and reciprocation of cooperation. For those concerned primarily with maximizing their own payoffs, the matrix may have put the global interdependence of the two players into view.

In the next study, we used a related game to explore further how other aspects of social context may interact with individual power motives in predicting aggression.

3 Effects of Power Motives and Interaction Partners

The Intergroup Prisoner’s Dilemma with Intragroup Power Dynamics (IPD^2) is a new paradigm for studying behavior in conflict situations. IPD^2 adds the concept of intragroup power to an intergroup version of the standard IPD. The game and experimental results are described in greater detail by Juvina et al. [13]. This addition is intended to bring our research one step closer to natural settings where individual tendencies toward aggression can be heightened or diminished in reaction to the behavior of others.

Traditionally, game theory studies the asymmetry in information available to individual decision-makers. Yet, a real-world conflict could be asymmetric in many other ways [14]. For example, power imbalances are known to be involved in radicalization of conflicts [15, 16]. When one faction’s power is disproportionately higher than that of a competing faction, and members of the latter see no realistic way of improving their standing through accepted political means, using extreme measures may become a rational choice [17].

Our study of behavior in IPD^2 immerses participants in situations where both teammates and opponents use a broad range of strategies, to study the impact of other players' behavior on the radicalization of a human player.

3.1 Overview of IPD^2

In IPD^2, two groups of two players interact. In each round, both players in a group "vote" whether to cooperate or defect, but the more powerful member of each group determines the group's decision with respect to the other group. Although these roles can shift throughout the game, we refer to the group member with greater power at any point in time as the "majority" and the less powerful member as the "minority."

Group payoffs are then distributed proportionally to power (with the player in majority receiving a larger share of either gains or losses), and power is shifted according to decision quality: If both group members made the same decision, power only changes by a small random amount; If they made opposite decisions and group payoff was positive, power increases for the player in majority (who enacted the successful decision); If they made opposite decisions and group payoff was negative, power increases for the player in minority (who opposed the unsuccessful decision). Power within a group is a constant quantity, such that when it increases for one player it decreases for the other accordingly.

3.2 Method

We recruited 130 participants (age M = 23.8 years, SD = 2.96 years; 62% male) to play IPD^2 in various conditions. Each participant played with a random set of pre-programmed strategies in the roles of group mate and members of the opposing group. These included the straightforward strategies: always cooperate and always defect, the classic strategy tit-for-tat (which repeats its opponent's most recent move), plus two more elaborate strategies that change behavior depending on whether they are in power. These latter strategies, labeled seek-power and exploit, are described below.

Seek-power plays like tit-for-tat when in the majority. When in the minority, it plays C or D depending on the outcome of the intergroup PD in the previous round. The logic of this choice is that the minority player tries to gain power by sharing credit for positive outcomes and avoiding blame for negative ones. In addition, this strategy makes an assumption with regard to other players' most likely moves, namely that majority players will repeat their moves from the previous round. If the previous outcome was symmetric (CC or DD), seek-power plays C. Under the assumption of stability (i.e., the previous move is repeated), C is guaranteed to not decrease power (except for random fluctuations). If the previous outcome was asymmetric (CD or DC), seek-power plays D, which is guaranteed not to lose power. Seek-power was designed as a 'best response' strategy that deals with predictable opponents. It was expected to play reasonably well and introduce a challenge for human players.

Exploit plays another classic strategy, 'win-stay, lose-shift', when in the majority and plays seek-power when in the minority. (win and lose refer to positive and negative payoffs, respectively.) 'Win-stay, lose-shift' (also known as Pavlov) is another very effective strategy that is frequently used against humans and other computer strategies in experiments. It has two important advantages over tit-for-tat: it recovers from occasional mistakes and exploits unconditional cooperators [18]. By combining Pavlov in the majority with seek-power in the minority, exploit was intended to be a simple yet effective strategy in IPD^2.

Each human participant was matched with each computer strategy twice within the same group, along with different pairs of the other strategies on the opposing team. Selection was balanced to ensure that each strategy appeared equally often both in the same group and in the opposite group. As a result, 10 game types were constructed. A Latin square design was used to counterbalance the order of game types. Thus, all participants played 10 games in one of 10 orderings. Each game was played for 50 rounds, such that each participant played a total of 500 rounds.

By using pre-programmed agents, we manipulated the behavior of the players that are matched with the human players. This design allowed us to analyze which experimental conditions triggered or facilitated extreme aggressive behavior in participants. We used a qualitative analysis to describe the aggressive tendencies of individuals that arose only when playing with certain combinations of strategies.

3.3 Results

In this study, we focus on the extreme aggressive behavior of the participants and specifically on how this behavior was influenced by the various conditions of the game. An aggressive action is defined as a repetition of D after a negative outcome resulting from that choice (e.g., a loss in both payoff and power). A player displayed extreme aggression in a particular game if their total number of aggressive actions was higher than three standard deviations above the mean across all participants in that game type. The following analysis is geared toward identifying the individuals who manifested extreme aggression and understanding what conditions are more conducive to radicalization of behavior. We will discuss only the game types that triggered extreme aggression in at least one participant.

In game type 2, four participants showed extreme aggression. Participants were grouped with exploit, and played against always-cooperate and tit-for-tat. A qualitative analysis of the game dynamics revealed that participants exhibiting extreme aggression defected most of the time, being occasionally rewarded with large increments in payoff and power when always-cooperate was in majority. By constantly defecting, these players managed to secure levels of payoff and power comparable with those of the other participants. Thus, although this behavior was extreme, it was rational from the perspective of a self-interested participant who chose to exploit the maximum predictability and altruism of always-cooperate.

In game type 3, four other participants manifested extreme aggression. Participants were grouped with always-cooperate, and played against seek-power and tit-for-tat. Looking at the game dynamics, it appeared that extreme participants sought power at the expense of payoff. When in majority they mostly defected and occasionally cooperated but only as much as needed to maintain power. They did not engage in long-lasting mutual cooperation with the opposite group, which would have secured a moderate level of payoff but not increased power (because their group mate would always cooperate as well). When in minority, they quickly learned that cooperation could bring them back to power. Compared to other participants, those showing extreme aggression obtained significantly lower payoff ($t(128) = -5.64$, $p < 0.001$), but non-significantly higher power ($t(128) = 1.58$, $p = 0.12$). Confirming that these four individuals were indeed driven by power, their power level in all games was marginally higher than that of the other participants ($t(128) = 1.93$, $p = 0.06$), while their payoff levels were not significantly different from those of other participants.

In game type 6, three participants manifested extreme aggression. One of these three participants was also an extremely aggressive player in game type 2. Participants were grouped with tit-for-tat, and played against always-defect and always-cooperate. Analysis of game dynamics revealed that the three extreme players defected most of the time and this behavior was occasionally rewarded when always-cooperate came to power. This behavior appears similar to that employed by the extreme players in game type 2. Here, the three participants showing extreme aggression obtained significantly more power than others ($t(128) = 2.64$, $p = 0.009$) and non-significantly lower payoff ($t(128) = -1.6$, $p = 0.11$). Increases in power were a byproduct of self-interested behavior of the human player combined with the ineffective minority play of tit-for-tat, which never got to power when grouped with these extreme players. This interpretation is supported by the observation that these three participants did not obtain higher levels of power in the other games.

In game type 9, only one participant manifested extreme aggression. Participants were grouped with always-defect, and played against tit-for-tat and exploit. The extreme participant started the game by cooperating, but switched to consistent defection at round 3. As a result, this player never came into power, but occasionally benefitted from the gains in payoff brought by the defecting majority. It appeared that this player's extreme behavior was an imitation of the group mate always-defect, combined with an acceptance of free-rider status. The payoff and power of this extreme participant were not significantly different from those of the others players.

3.4 Discussion

Participants were exposed to various game types defined by the behavior of pre-programmed agents. Four of these conditions triggered extreme aggressive behavior in different individuals, with the only exception of a participant who behaved aggressively in two conditions. This leads us to conclude that neither characteristics pertaining to the individual nor those related to the context are solely responsible; it is their combined effects that determined extreme aggressive behavior.

Individual characteristics can certainly be inferred. The individuals exhibiting extreme aggression in game types 2 and 6 appeared to be narrowly self-interested in terms of payoff and unable to understand the logic of a non-zero-sum game. They neglected the opportunity to engage in mutual cooperation, which would have yielded even higher payoffs. Those with extreme behavior in game type 3 appeared to be driven exclusively by power. Their behavior was counterproductive in terms of payoff, but fruitful in attaining majority status. Finally, the extreme player in game type 9 seemed to imitate a group mate's extreme aggression, and shared in the occasional benefits for doing so (at the expense of the opposite group).

However, participants demonstrated extreme aggression in only a limited set of conditions, suggesting that contextual factors played an important role. In game types 2 and 6, a reliably altruistic player (always-cooperate) in the opposite group ensured that a player's defection was sometimes rewarded with large payoffs. In game type 3, having always-cooperate in one's group made it difficult for people to increase power by cooperating (since power changes only a small random amount if group members make the same decision). Thus, defection was the preferred alternative for someone strongly motivated by power. In game type 9, having always-defect in power within one's group made it easy to maintain a decent level of payoff without much effort. Always agreeing with the majority decision to defect maintained power close to that of the majority, which secured an almost equal share of group payoffs.

4 Conclusions

Much prior research on predicting aggression in conflict has emphasized the impact of either stable individual traits [19, 20] or contextual factors [21, 22] in isolation. However, few researchers have considered the interaction between individual proclivities and features of the decision environment. Here, we took this novel approach in two variations of the PD.

In our first study, we tested the distinct influence of individual background variables on defection rates in the IPD across conditions differing in information. In the Description condition, participants had continuous access to a complete payoff matrix mapping players' actions to outcomes; In the Experience condition, participants only discovered these mappings over time. From a rational standpoint, these two conditions should have produced no differences in behavior, and yet we found that those with higher interpersonal trust and maximizing tendency scores more often chose defection as a means to achieve their goals in the Experience condition. This suggests that learning interdependence only as outcomes are experienced over time, as is typical in the real world, may hinder realizations that cooperation will build the trust of others and even increase one's own well-being in the long run.

In the second study, we assessed the tendencies toward extreme levels of aggression in the IPD^2. Our experimental design provided participants with power motives and a range of interaction partners, also common in actual conflict. We observed that few players ever behaved in an extreme way, and even those who appeared extreme in some game types were not consistently so across other games. This reinforces our conclusion that it is neither individual nor situational factors alone that induce extreme actions, but rather a confluence of the two.

References

1. Huesmann, L.R.: An Information Processing Model for the Development of Aggression. Aggressive Behav. 14, 13–24 (1988)
2. Rapoport, A., Chammah, A.M.: Prisoner's Dilemma: A Study in Conflict and Cooperation. University of Michigan Press, Ann Arbor (1965)
3. Nash, J.: Equilibrium Points in N-person Games. P. Natl. Acad. Sci. USA 36, 48–49 (1950)
4. Rabin, M.: Incorporating Fairness into Game Theory and Economics. Am. Econ. Rev. 83, 1281–1302 (1993)

5. Kreps, D.M., Wilson, R.: Reputation and Imperfect Information. J. Econ. Theory 27, 253–279 (1982)
6. Martin, J.M., Gonzalez, C., Juvina, I., Lebiere, C.: The Description-Experience Chasm in Social Interaction (under review)
7. Barron, G., Erev, I.: Small Feedback-based Decisions and their Limited Correspondence to Description-based Decisions. J. Behav. Decis. Making 16, 215–233 (2003)
8. Yamagishi, T.: The Provision of a Sanctioning System as a Public Good. J. Pers. Soc. Psychol. 51, 110–116 (1986)
9. Diab, D.L., Gillespie, M.A., Highhouse, S.: Are Maximizers Really Unhappy? The Measurement of Maximizing Tendency. Judgm. Decis. Mak. 3, 364–370 (2008)
10. Jones, E.E., Harris, V.A.: The Attribution of Attitudes. J. Exp. Soc. Psychol. 3, 1–24 (1967)
11. Kahneman, D., Tversky, A.: Prospect Theory: An Analysis of Decision under Risk. Econometrica 47, 263–291 (1979)
12. Premack, D., Woodruff, G.: Does the Chimpanzee have a Theory of Mind? Behav. Brain Sci. 1, 515–526 (1978)
13. Juvina, I., Lebiere, C., Martin, J.M., Gonzalez, C.: Intergroup Prisoner’s Dilemma with Intragroup Power Dynamics. Games: Special issue on Predicting Behaviour in Games 2, 21–55 (2011)
14. Sandler, T., Arce, D.G.: Terrorism: A Game-Theoretic Approach. In: Sandler, T., Hartley, K. (eds.) Handbook of Defense Economics, pp. 775–813. Elsevier, Amsterdam (2007)
15. Crenshaw, M.: The Causes of Terrorism. Comp. Polit. 13, 379–399 (1981)
16. Crenshaw, M.: Decisions to Use Terrorism: Psychological Constraints on Instrumental Reasoning. Int. Soc. Movements Res. 4, 29–42 (1992)
17. Atran, S.: Genesis of Suicide Terrorism. Science 299, 1534–1539 (2003)
18. Nowak, M., Sigmund, K.: A Strategy of Win-Stay, Lose-Shift that Outperforms Tit-for-Tat in the Prisoner’s Dilemma Game. Nature 364, 56–58 (1993)
19. Huesmann, L.R., Eron, L.D., Lefkowitz, M.M., Walder, L.O.: Stability of Aggression Over Time and Generations. Dev. Psychol. 20, 1120–1134 (1984)
20. Dodge, K.A., Laird, R., Lochman, J.E., Zelli, A.: Multidimensional Latent-Construct Analysis of Children’s Social Information Processing Patterns: Correlations With Aggressive Behavior Problems. Psychol. Assessment 14, 60–73 (2002)
21. Kupersmidt, J.B., Griesler, P.C., DeRosier, M.E., Patterson, C.J., Davis, P.W.: Childhood Aggression and Peer Relations in the Context of Family and Neighborhood Factors. Child Dev. 66, 360–375 (1995)
22. Price, J.M., Dodge, K.A.: Reactive and Proactive Aggression in Childhood: Relations to Peer Status and Social Context Dimensions. J. Abnorm. Child Psych. 17, 455–471 (1989)

Scent Trails: Countering Terrorism through Informed Surveillance

Alex Sandham[1], Tom Ormerod[1], Coral Dando[1], Ray Bull[2],
Mike Jackson[3], and James Goulding[3]

[1] SCORPIO Centre, Psychology Department, Lancaster University, UK
[2] Psychology Department, Leicester University, UK
[3] Centre for Geospatial Science, Nottingham University, UK
{a.sandham,t.ormerod,c.dando}@lancaster.ac.uk,
ray.bull@le.ac.uk,
{mike.jackson,james.goulding}@nottingham.ac.uk

Abstract. This paper reports the DScenT (Detecting Scent Trails) project, which brought together technologists and behavioural scientists to design and evaluate novel methods for countering terrorism in public places. Through a mixture of prototyping and empirical evaluations, we developed and assessed an immersive environment for detecting and investigating deceptive behaviours indicative of terrorist activities. The environment comprised a location-based game called Cutting Corners. The game was used in field trials to test the efficacy of different methods for collecting and using evidence during investigative interviews with mock terrorist suspects and to examine effects of play on attitudes towards surveillance and counter-terrorism.

Keywords: Countering Terrorism; Pervasive Games, Location-based Games, GIS, Detecting Deception, Suspect Interviewing.

1 Introduction

At a time of raised terrorist threat, there is a need for research that assists security and police services in protecting the public and key assets. The United Kingdom's Strategy for countering terrorism [1] describes how the UK aims to reduce risk of terrorism with an emphasis on the four "p's" of prevention, pursuit, protection and preparation. The DScenT project addressed the ‘Pursuit’ theme. Operations to intervene with suspects are high risk because inappropriate surveillance, interview or arrest may have damaging political, public relations and intelligence effects. In addition to better tracking information, the security and police services need to have confidence that operations yield evidence which can demonstrate conclusively that a deceptive activity was in the process of being planned or executed when an operation took place. By sensitive application of human-centred HCI principles to technology design, we aimed to provide tools for effective evidence gathering and interpretation.

1.1 Scent Trails for Evidence Gathering and Use

The thesis of the DScenT project was that collective movements and communications of persons working together in planning and executing a terrorist event may provide

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 452–460, 2011.

(i) an indication of intent prior to its conclusion and therefore assist early intervention, (ii) a reduction in suspect false positives enabling more targeted surveillance, and (iii) an ability to schedule surveillance or carry-out disruptions based on suspect patterns of behaviours. The concept of a scent trail was developed where, by measuring movement, communications and behaviours of individuals working as a team, a rich and integrated time-, interaction- and location-based history could be constructed. The challenge was to determine whether scent trails can be analyzed to reveal anomalous behaviours that differentiate between activities unrelated to serious crime and deceptive activities with a terrorist intent.

One use for scent trails is issuing 'challenges' to suspects, both in real-time while they are being monitored and during interviews after an arrest. A challenge might consist of presenting a suspect with scent trail information (e.g., two individuals who deny knowledge of each other being in the same unusual location at the same time). Scent trail challenges can yield two benefits. First, the challenge might undermine a suspect's account during an interview. Second, being presented with incriminating data might change a suspect's behaviour, leading to the aborting of an attack.

These hypotheses follow from studies of police interviewing techniques, which show that interviewers do not challenge suspects' accounts. Instead, a typical tactic is to reveal evidence early in an interview to 'invite' the suspect to confess, which only succeeds when evidence is unequivocal [2]. The *PEACE* approach was designed to overcome this problem, by eliciting an account from the interviewee, and then raising inconsistencies between their account and evidence. In this way, interviewers avoid misleading cues to lying (e.g., gaze aversion or fidgeting), and focus instead on the content of interviews, which provides reliable cues to deception [3]. However, the interviewer needs worthwhile evidence to point out errors or omissions. High quality evidence is critical in the context of counter-terrorism, where even holding an interview can have negative impacts if it cannot be clearly justified.

1.2 Using Location-Based Games to Explore Scent Trails

Terrorist activities and investigative practices used to curtail them are too sensitive to be studied directly. Yet, given the continual evolution of terrorist methods, to limit research to known event types (e.g., the 7/7 bombings) is counter-productive. We explored scent-trails in a non-sensitive simulation that provides an analogy to deceptive activities, and which mixes "minor" civil deceptions with deception linked to a terrorist event to simulate the complexity of real crime detection.

With few exceptions (e.g., cyber-terrorism) terrorist activity has a spatial or geographic context. Spatial analytical techniques to anticipate areas of crime or terrorism, and to gather intelligence and intercept such activity is a growing area of research [4]. Technologies such as ubiquitous indoor-outdoor positioning, remote sensing and surveillance, high-bandwidth mobile communications and spatial search, linked to semantic web and location-based services, are developing rapidly. These provide powerful analytical tools which can help interception, event response and detection of criminal/terrorist activity and individuals or networks.

Positioning data can be analyzed to find behaviour patterns and interactions, whose outcomes constitute location-based intelligence. It is possible to deduce information such as where the participant is located in time and the speed at which they are travelling, whether they would have had signal coverage, and so on. The speed at which someone is travelling can indicate what transport they are using. How long they spend at a location suggests activities they engage in [5].

1.3 Pervasive and Location Based Games

Including physical activity and social interaction using ubiquitous technologies allows computer games to be integrated with physical and social aspects of the real world to create *Location-based games (LBGs:* Benford, Magerkurth, & Ljungstrand, 2005). LBGs treat the real world as a game board where the players act as interactive game pieces [6]. Mobile technologies allow tracking of players' positions as they interact. LBGs provide research opportunities to explore challenges of new technologies. For example, games such as *Can You See Me Now* [7], *Uncle Roy All Around You* [8], and *Feeding Yoshi* [9] use a mix of PDAs with GPS and WiFi, while games like *Botfighters* [10], and *The Day of the Figurines* [11] use mobile phones and SMS. WiFi black spots and GPS inaccuracies create periods of uncertainty during play, especially in urban environments which players learn to work around [12].

Serious games are concerned with purposes beyond the gaming experience itself. For example, pedagogy can be combined with entertainment to engage students [13]. Serious games have also been used to simulate real life situations that are costly or impossible to create in any other way. They provide a safe environment where training or exploration can be gained at low cost. Serious games have been used for military training [14], health training [15], therapy and rehabilitation [16], education [17], and emergency response [18]. LBGs provide a new class of serious games, offering controlled 'experimental' environments where psychological theories about human behaviour be tested [19]. LBGs may provide training tools for stressful and dangerous environments as they provide a safe and realistic training ground. For example, *Rogue Signals* mirrors emergency fire response teams [20].

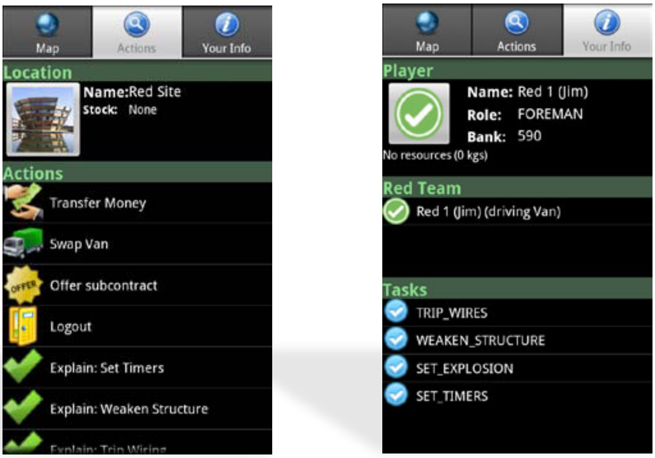
2 The Cutting Corners Game

Cutting Corners is an LBG research tool to understand how best to reveal and challenge deceptive behaviour by combining location sensitive technology with forensic psychology techniques for detecting deception. The game is a collaborative strategic task-orientated game where teams are assigned objectives and team members assume different roles. The game design allows players to develop their own strategy with the opportunity to cheat, which promotes various levels of deceptive behaviour. This focused design strategy provides a vehicle to explore more serious objectives within a controlled environment. Cutting Corners provides a safe non-sensitive environment that simulates the planning of a terrorist attack which is unfeasible to explore any other way. A unique feature of Cutting Corners is that it provides, for the first time, an environment for testing methods of detecting deception in which participants create their own deceptions rather than being presented with a set of statements or activities to lie about.

The research utilises GPS-enabled mobile communications to collect *scent trials* (i.e., records of an individual's passage and activities across time through a physical space) as people play the game. Players' behaviours are monitored by integrating their movements via GPS, their patterns of communication via mobile phones and transactions made during the game, which are all collated to form unique player scent trails. Scent trails provide detailed accounts of game play activities, allowing investigators to make predictions and plan for interventions if deemed necessary. These scent trails are then further used as evidence during post-game interviews to determine the veracity of players' accounts of their gaming. Each thread within a scent trail is evaluated to understand what combinations of technology have the most impact in determining deceptive behaviour. Using an ethnographic research method, the players' experiences can be analysed, to provide a valuable resource to investigate player behaviours such as team collaboration and decision making.

2.1 The Structure of the Game

Cutting Corners involves four teams of three players simulating the role of building contractors racing to build an event stage. The first team to complete wins the game. Teams are told to 'bend the rules' to finish quicker. One team is briefed to undertake a terrorist attack while masquerading as builders. The game area consists of an event site, virtual checkpoints, game start and four shops where players purchase tokens representing needed items. The game winners are the team who completes their tasks first by assembling all the necessary items at the event site. Prior to the game, each team is briefed, where the game rules and interface (see Fig. 1) are described. Another group has the role of investigator, trying to detect the terrorist team. Investigators are presented with limited information to simulate real surveillance gaps and may carry out '*security checks'* in which they stop players and inspect their purchases. Scent trails provide evidence from which the investigator collates a case file used to identify the terrorist team. The game is designed so players can develop a strategy that enhances deceptive behaviour yet avoid detection.



Name:Red Site
Stock: None

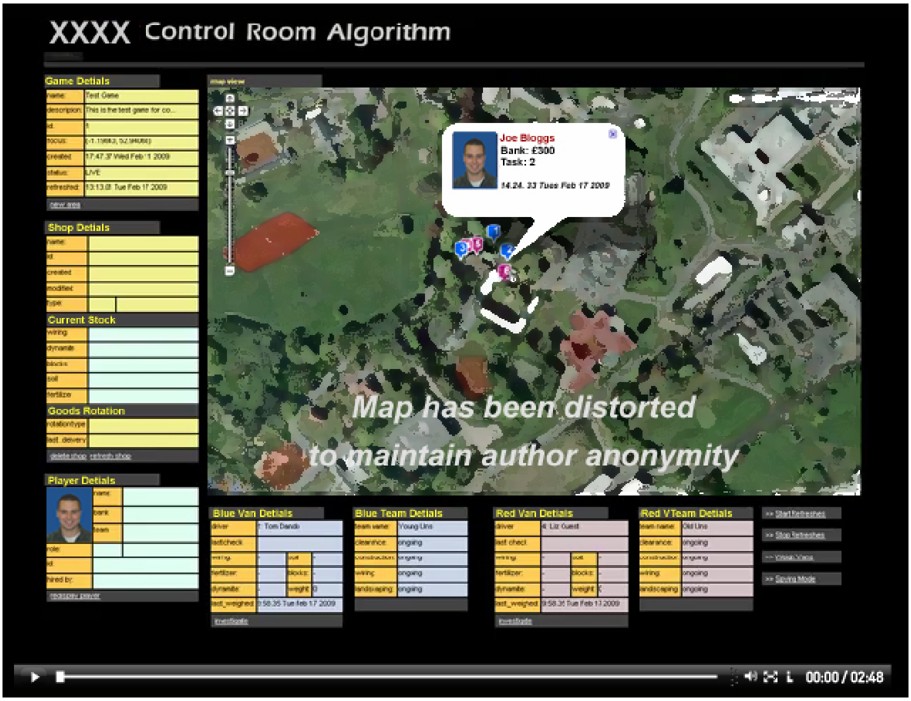
Name: Red 1 (Jim)
Role: FOREMAN
Bank: 590
No resources (0 kgs)

Red 1 (Jim) (driving Van)

Fig. 1. The Cutting Corners mobile phone application. The 'Map' tab (left) presents a topographical overview of key locations, 'Actions' (centre) shows players current location and 'Your info' (right) shows current status of the player's team.

2.2 The Game Technology

Each player is supplied with a 3G mobile phone with built in GPS and Wi-Fi support which runs a bespoke games software interface implemented in the Android operating system. The software uses the phone web browser to communicate to the main server which continually updates the main database throughout the game. It relays the GPS coordinates back to the main server and controls all game play activities. Team members can communicate with each other via their phone. The whole game is monitored by a control room where a central server tracks the player's movements in real time through their mobile device. This is represented on a 3D map displayed in the control room which shows all the players game play activities as and when they occur, as shown in Fig. 2. Each game is recorded for later play back where different levels of information can be replayed to determine the optimum scent trail which provides enough information without overloading the investigators.



Joe Bloggs
Bank: £300
Task: 2
14.24.33 Tues Feb 17 2009

Game Detials	
name:	Test Game
description:	This is the test game for co...
id:	1
focus:	(-1.15863, 52.94586)
created:	17:47:37 Wed Feb 11 2009
status:	LIVE
refreshed:	13:13:01 Tue Feb 17 2009

Shop Detials		
name:		
id:		
created:		
modified:		
type:		

Current Stock	
wiring:	
dynamite:	
blocks:	
soil:	
fertilize:	

Goods Rotation	
rotationtype:	
last_delivery:	

Player Detials	
name:	
bank:	
team:	
role:	
id:	
hired by:	

Blue Van Detials			
driver:	1: Tom Dando		
lastcheck:			
wiring:	-	soil:	-
fertilizer:	-	blocks:	-
dynamite:	-	weight:	0
last_weighed:	9:58:35 Tue Feb 17 2009		

Blue Team Detials	
team name:	Young Uns
clearance:	ongoing
construction:	ongoing
wiring:	ongoing
landscaping:	ongoing

Red Van Detials			
driver:	4: Liz Guest		
last check:			
wiring:	-	soil:	-
fertilizer:	-	blocks:	-
dynamite:	-	weight:	0
last_weighed:	9:58:35 Tue Feb 17 2009		

Red VTeam Detials	
team name:	Old Uns
clearance:	ongoing
construction:	ongoing
wiring:	ongoing
landscaping:	ongoing

Fig. 2. The Control Room interface showing the game play activities in real time

3 Empirical Study

3.1 Experiment to Test Suspect Interview Techniques

To date, we have used the Cutting Corners game to compare different investigative interviewing techniques, involving early, late and gradual revelation of scent trail evidence during suspect interviews. In empirical trials, scent trail information was presented to experienced professional police investigators. They used the scent trails during games to detect and intercept players who appeared to be undertaking potentially terrorist-related activities (e.g., simulating the detection of a 'dry run') and after the game in investigating suspicions developed by observing game play via post-game interviews: investigators interviewed each player in order to determine whether they were legitimate or deceptive, with rewards received by players who remained undetected. A limited description of the study is given here to illustrate the value of LBGs in crime detection research.

3.2 Method

Participants. 96 graduates aged between 18 and 45 were recruited from Internet bulletin boards and personal contacts. They were paid £30 to take part and were instructed that they would receive additional performance related payments (£10 each for the winning team, plus £5 for remaining undetected in the interview phase). To ensure parity, all participants received the additional payments regardless of performance.

Design. Participants were assigned to either terrorist or builder groups, and each group was further sub-divided into participants receiving early, late or incremental scent trail revelation during post-game interviews.

Procedure. Participants received pre-game instruction via video, and were then taken to the game start point. Teams were given 20 minutes to prepare a strategy by which they would aim to maximise their chances of winning the game. On a signal from the experimenter, the game commenced and participants set off to visit their relevant game locations, meeting up at intervals as required by game rules to transfer items purchased in different locations.

A set of rules embodied in the game software determined when and where the players could make and transfer purchases (e.g., there was only one virtual 'truck' in which to carry purchases, which was a software applet available on only one player's phone at any one time, but which could be transferred between co-located players). These rules were designed a) to add a problem-solving and planning component to game play, b) to encourage the teams to strategize, and c) to enable opportunities for deceptive activities to arise. The phone interface described the purchase and delivery requirements for each of four tasks that each team had to complete in sequence (the terrorist team(s) simulating parts of the builders' tasks as cover while carrying out a smaller set of terrorist-specific tasks). The game took approximately one hour to complete. At the end of the game, players completed a questionnaire, in which they evaluated the game for interest, difficulty and enjoyment.

In addition to game rating, players were also given a post-game questionnaire that explored attitudes to surveillance. Players gave ratings on a Likert scale (1= strongly negative, 7 = strongly positive) to questions focussed upon attitudes towards use of surveillance for counter-terrorism and commerce (e.g., purchase monitoring), and to the perceived level of terrorist attack threat. Player ratings were compared to ratings given by a sample of 40 members of the general public and 14 police officers.

During the game, investigators monitored players' scent trails to detect anomalies indicative of terrorist activity. They took notes on each player's activities as revealed by the scent trails to serve as evidence for use in post-game interviews. At the end of the game, investigators made a judgement about each player (they were naïve to team membership) as to whether they believed them to be a builder or terrorist. Then each player was interviewed for approximately 15 minutes using the PEACE approach to interviewing (free recall followed by a questioning phase), in which scent trail evidence was revealed either immediately, at the end of the recall phase, or incrementally during the recall.

3.3 Results and Discussion

Detection accuracy at the end of the game (pre-interview) was 67% (terrorists 64%, builders 70%). While no comparison data exist from other studies of on-going deception detection rates, this result compares favourably with mean rates in published laboratory studies of deception detection *after* interview with suspects, which have an average accuracy rate of 57% [21]. Thus, scent trails seem to perform reasonably well as a cue to deception detection judgements. Moreover, scent trails seem particularly to support the identification of innocent players, a bias that is not evident in other studies.

Post interview accuracy was 80%, suggesting scent trails provide investigators with powerful evidence source for detecting and proving deception. The accuracy rate for incremental revelation of evidence was 91%, compared with 80% for late revelation and 69% for early revelation. Thus, Cutting Corners appears to provide a usable test environment for evaluating different approaches to detection of deception.

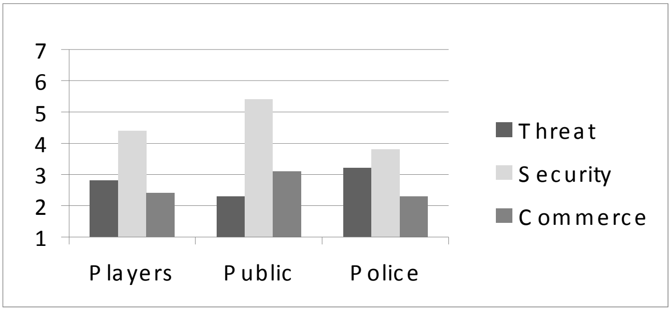


Fig. 3. Ratings of attitudes about surveillance by players, general public and police officers

The results of the attitudinal questionnaire are illustrated in Fig. 3. Overall, the general public are most accepting of surveillance for security reasons, and surprisingly the police are least accepting, despite the police seeing the threat level as higher. The effect of immersion in the cutting corners game seems to be to make players' judgements close to police officers. We suggest that playing the game has a dual effect: players gain a heightened awareness of potential threats, but at the same time become more concerned about the role that surveillance plays in countering that threat. This may occur either because players gain a greater understanding of how surveillance can be used to advantage by laying false scent trails to appear less suspicious, or because the experience of being closely monitored is uncomfortable.

4 Conclusions

The outcomes of the research demonstrate the potential for designing and using location-based game experiences as an empirical test-bed for research that cannot otherwise be conducted in realistic environments. The game itself turns out to be engaging for players, whose levels of enjoyment and motivation remain high throughout, even though game play (including instruction and post-game interview) lasts over two hours. The game offers the first empirical environment for forensic psychologists to test methods of deception detection in which individuals construct their own deceptive activities that are not scripted by the experimenter. It also provides a realistic environment in which to examine how the planning and reactive strategies of deceptive and non-deceptive individuals and groups differ. The player interface demonstrates how geospatial, communications and commerce channels can be integrated to allow real-time game play across difficult terrains. The investigator interface explores how different forms of surveillance signal can be integrated within a single 'scent trail' representation to support complex investigative decision-making. Finally, heuristic algorithms can be layered onto scent trail representations to provide predictive inferences concerning player activities (e.g., inferring the members of a terrorist group, and what their individual roles within the group may be).

Acknowledgements. The DScenT project was funded by the EPSRC/ESRC/AHRC Countering Terrorism in Public Places Initiative, EP/F008686/1, EP/F006500/1, EP/F008562/1, EP/F008600/1, and EP/F014122/1, with co-funding from the UK Centre for the Protection of National Infrastructure. The authors gratefully acknowledge support of the other team members and Jenny Hart.

References

1. UK Government. Countering International Terrorism: Cm 6888 (July 2006)
2. Milne, R., Bull, R.: Investigative interviewing Psychology & practice. Wiley, Chichester (1999)
3. Vrij, A.: Detecting lies and deceit: The psychology of lying and it's implications for professional practise. Wiley, Chichester (2000)
4. Taylor, P.J., Snook, B., Bennell, C.: The bounds of cognitive heuristic performance on the geographic profiling task. Applied Cognitive Psychology 23, 410–430 (2009)
5. Liao, L., Fox, D., Kautz, H.: Learning and inferring transportation routines. In: Proc. National Conference on Artificial Intelligence, AAAI (2004)
6. Nicklas, D., Pfisterer, C., Mitschang, B.: Towards Location-Based Games. In: Sing, L.W., et al. (eds.) Proceedings. International Conference of Applications and Development of Computer Games in the 21st Century (ADCOG 21) (2001)

7. Anastasi, R., Tandavanitj, N., Flintham, M., Crabtree, A., Adams, M., Row-Farr, J., Iddon, J., Benford, S., Hemmings, T., Izadi, S., Taylor, I.: Can You See Me Now? A Citywide Mixed-Reality Gaming Experience. In: Proceedings of Ubi-Comp 2002, Gothenburg, Sweden (2002)
8. Benford, S., Magerkurth, C., Ljungstrand, P.: Bridging the Physical and Digital in Pervasive. Gaming, Communications of the ACM 48(3), 54–57 (2005)
9. Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown, B., Rowland, D., Benford, S., Capra, M., Hampshire, A.: Interweaving Mobile Games with Everyday Life. In: Conference on Human Factors in Computing Systems, Montreal, Canada, April 2006, pp. 417–426 (2006)
10. Sotamaa, O.: All The World's A Botfighter Stage: Notes on Location-based Multi-User Gaming. In: Computer Games and Digital Cultures Conference Proceedings, Tampare (2002)
11. Flintham, M., Smith, K., Benford, S., Capra, M., Green, J., Greenhalgh, C., Wright, M., Adams, M., Tandavanitj, N., Row-Farr, J., Lindt, I.: Day of the Figurines: A Slow Narrative-Driven Game for Mobile Phones Using Text Messaging. In: Proceedings of Pergames 2007,Salzburg (2007)
12. Chalmers, M., Galani, A.: Seamful interweaving: heterogeneity in the theory and design of interactive systems. In: Symposium on Designing Interactive Systems, August 1–4, pp. 243–252. Cambridge, Massachusetts (2004)
13. Allen, L., Seeney, M., Boyle, L., Hancock, F.: The Implementation of Team Based Assessment In Serious Games. Paper presented at the Proceedings of the IEEE Virtual Worlds for Serious Applications, Coventry, UK (2009)
14. Numrich, S.K.: Culture, Models, and Games: Incorporating Warfare's Human Dimension. IEEE Intelligent Systems, 58–61 (July/August 2008)
15. Sawyer, B.: Cells to Cell Processors: The Integration of Health and Video Games. In: IEEE Computer Graphics and Applications, November/December 2008, pp. 83–85 (2008)
16. Burke, J.W., McNeill, M.D.J., Charles, D.K., Morrow, P.J., Crosbie, J.H., Donough, S.M.: Serious Games for Upper Limb Rehabilitation Following Stroke. In: Conference in Games and Virtual Worlds for Serious Applications, IEEE VS-Games 2009, Coventry, UK, pp. 103–110 (2009)
17. Ardito, C., Buono, P., Costabile, M.F., Lanzilotti, R., Pederson, T., Piccinno, A.: Experiencing the Past through the Senses: An M-Learning Game at Archaeological Parks. In: IEEE MultiMedia, October–December 2008, pp. 76–81 (2008)
18. Chittaro, L., Ranon, L.: Serious games for training occupants of a building in personal fire safety skills. In: Conference in Games and Virtual Worlds for Serious Applications, IEEE VS-Games 2009, Coventry, UK, pp. 76–83 (2009)
19. Girardin, F., Nova, N.: Getting real with ubiquitous computing: The impact of discrepancies on collaboration, eMinds No 1 (2006), `http://www.hci.uniovi.es/en-eminds.htm`
20. Toups, Z.O., Kerne, A.: Coordination in Firefighting Practise: Design Implications for Teaching Fire Emergency Responders. In: Conference on Human Factors in Computing Systems, San Jose, CA (2007)
21. Vrij, A.: Detecting deception. Wiley, Chichester (2008)

Using Behavioral Measures to Assess Counter-Terrorism Training in the Field

V. Alan Spiker[1] and Joan H. Johnston[2]

[1] Anacapa Sciences, Inc.
Santa Barbara, California
[2] Naval Air Warfare Center Training Systems Division
Orlando, Florida

Abstract. Development of behavioral pattern recognition and analysis skills is an essential element of counter-terrorism training, particularly in the field. Three classes of behavioral measures were collected in an assessment of skill acquisition during a US Joint Forces Command (JFCOM)-sponsored course consisting of combat tracking and combat profiling segments. These included situational judgment tests, structured behavioral observation checklists, and qualitative assessments of the emergence of specific knowledge-skills-attitudes over the course of training. Evidence was present in all three types of measures to indicate that behavioral pattern recognition and analysis skills were successfully acquired by most students (a mix of Army and civilian law enforcement personnel). The paper describes both the types of skills acquired and the statistical evidence that supports their acquisition over the course of field training. Implications for broader training of these critical skills are also discussed.

Keywords: **S**ituational judgment tests, behavioral observations, scenarios, knowledge-skills-attitudes, profiling, tracking.

1 Introduction

In 2008, the US Department of Defense placed Irregular Warfare (IW) on an equal footing with conventional warfare in future military planning and operations [1]. Among IW mission objectives are developing capabilities to address asymmetrical threats and the challenges they pose for counter-terrorism (CT). Whether practiced by military ground units or law enforcement personnel (e.g., Custom and Border Patrol), reading the human terrain – such as through behavioral pattern recognition and analysis (BPRA) skills – is an essential element of CT training. BPRA is a set of skills and techniques that a profiler uses (e.g., cues and indicators of behaviors) to spot people and events *before* the situation becomes lethal. Staying 'left of bang' by constructing these behavior profiles in a proactive fashion is now considered to be a protective element for small units, and is every bit as important as body armor and weaponry [2]. As a result, behavior profiling techniques have become a valuable addition to small unit tactics, techniques, and procedures (TTPs).

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 461–470, 2011.

Nowhere are these skills more highlighted than in Combat Hunter (CH) training conducted by the United States Marine Corps School of Infantry. A 10-day course, CH is taught in two segments, combat profiling and combat tracking, by subject matter experts (SMEs) in these respective fields. Each segment has a classroom academic portion and scenario-based field exercises where the skills are developed, applied, refined, and reinforced. Historically, it is hard to measure training development in the field since the observation conditions are difficult, curriculum is not always standardized, objectives are not always well-specified, and instructional delivery by instructors is inconsistent.

To document training acquisition, a wide array of behavioral methods and measures are needed. In this paper, we summarize the results of a field study where we observed development and acquisition of BPRA skills by military and law personnel as part of formal counter-terrorism training in a special offering of CH called Border Hunter. Specifically, we describe empirical results using three different methods of behavior measurement: situational judgment tests (SJTs), behavior observation checklists (BOCs), and knowledge-skill-attitude (KSA) profiles. This discussion addresses both the content of the skills measured and evidence for their acquisition. We conclude with a discussion of the implications for using behavioral measures to support CT training within a broader spectrum of training technologies.

2 Training Counter-Terrorism Skills

2.1 Combat Hunter

Instruments to collect behavioral measures were developed during the authors' naturalistic observation of CH training at the School of Infantry – West, Camp Pendleton. CH is a ten-day course split equally between Combat Tracking and Combat Profiling. Each class consists of approximately 40 students drawn from the same Marine regiment, though students typically come from different platoons and squads. Both segments are split into academic instruction and field scenario portions. In Combat Tracking, students receive academic instruction in the morning on the fundamentals of tracking (e.g., dynamics of footprints, maintaining track line, interpreting spoor), where the afternoons are devoted to utilizing this knowledge in the field where they track 'quarry' (role-playing instructors) as five-person tracking teams. During these tracking scenarios, students learn to read their enemies' spoor (i.e., footprints, human signs, environmental cues, slight ground disturbances). They are also taught to build social/biometric profiles of their quarry, anticipate their targets' actions by acquiring the mindset of the quarry, and apply TTPs to hunt down their targets. Combat tracking is a human-centric competency particularly useful in IW settings to support offensive operations, intelligence gathering, clandestine movement in hostile areas, and counterinsurgency operations. Over days, the field scenarios increase in complexity as the instructors add in such factors as more difficult terrain, more 'skilled' quarry (e.g., where the role players purposely try to cover their tracks), and more intricate team tracking maneuvers.

Combat Profiling is structured differently, where 3 days of academic instruction precede the field scenarios. Profiling is concerned with perceiving, analyzing, and articulating critical events within the human terrain. Its main goal is to identify pre-event indicators through human behavior ‘left of bang’, (i.e., before a destructive event occurs). It trains individuals to look for behaviors that are anomalous, beyond the baseline of a culture or location. Through combat profiling, warfighters and law enforcers learn to be more situationally aware and to accurately interpret subtle cues that forewarn a critical event. During classroom instruction, students are exposed to the basic concepts of profiling, such as fundamentals of optics, pattern recognition, reasoning by analogies, forming prototypes, ethical-moral decision-making, and the six domains of combat profiling (i.e., heuristics, geographics, proxemics, atmospherics, biometrics, kinesics). The practical application portion is conducted in the next two days where students are split into teams and man observation posts (OPs) and observe role-players engaged in varying types of behavior within a village mockup. They practice their profiling skills by observing, at a distance, instances of neutral and insurgent behavior in the context of increasingly challenging scenarios. Typically, 5 to 6 such scenarios are executed during these two days. The training culminates in a 4-hour long final exercise where all teams deploy as maneuver units into the village using their insights gained from the previous scenarios.

The authors observed two evolutions of the CH course; these observations were used to develop draft versions of the three types of behavioral measure instruments [3]. Besides observing students and instructors, the authors interviewed selected students and instructors for further information, to clarify points, and to extract the higher level skills that were being trained. Additional materials were obtained from the instructors that provided further information concerning the theoretical and practical underpinnings of combat profiling and tracking.

This repository of information was then used to create separate versions of an SJT for Combat Tracking and Combat Profiling, as well as a structured BOC for each segment. The SJTs were six-item tests that required participants to think about and decide among six possible response options for each briefly-described scenario. They follow the format typically recommended for industrial applications [4]. The BOCs were structured so that rater/observers would provide 3- or 5-point ratings on a set of basic and advanced profiling or tracking skills. In addition, a taxonomy of KSAs was created that would be applicable to both course segments. The taxonomy would be used retrospectively by observers, after the training day’s conclusions, as a way to categorize the free-form observations they made while observing students. Further details on each instrument will be discussed when the empirical results are presented.

2.2 Border Hunter

Border Hunter (BH) can be characterized as a one-off, ‘graduate-level’ version of CH [4] that also consists of combat tracking and profiling instruction. Twenty-one days long, BH was sponsored by USJFCOM and was conducted by Joint Task Force North at Fort Bliss, TX in April 2010. The 10-day Combat Tracking segment was taught by six highly experienced trackers with a combined experience of 180 years. Similarly, Combat Profiling was taught by an eight-person team of instructors with a collective experience in military and police work of more than 200 years.

Forty-three trainees received the BH training, comprising a mix of US Army, Border Patrol, and other law enforcement personnel. All were highly experienced with an average of 9 years in military/law enforcement. In addition, 22 soldiers from Fort Bliss were recruited as role players during the Combat Profiling field scenarios.

A 13-person research team was present to capture course content for subsequent packaging, as well as to conduct behavioral research and assess training effectiveness. This included both the academic instruction and field training aspects of the course. The focus of the present paper is with the latter, where on a given day, 3 to 6 individuals were available to collect behavioral data on student training performance. All were highly experienced researchers with advanced degrees.

Students were assigned to the same teams for all Tracking and Profiling field training scenarios (FTXs). Teams 1 and 2 were composed of students from the Custom and Border patrol agencies. Teams 3 and 4 were configured with Army personnel; students in Team 3 were less experienced than their Team 4 counterparts. Team 5 was a hybrid team, comprising a mix of Army and law enforcement personnel.

3 Situational Judgment Tests

3.1 Method

SJTs are low- to moderate-fidelity work sample simulations that assess preferences for appropriate behaviors in a work setting [5]. While SJTs have long been used in industrial settings for job selection and placement, their use as a source of proficiency data from field settings is less frequent. Because SJTs have moderate concurrent validity with performance [6], we elected to use them in this project as a way to assess degree of learning in the FTXs for both Tracking and Profiling.

SJT items ask respondents to assess the effectiveness of various response options. The scenarios are intentionally written so that not all situational cues are known, which increases dependence on one's judgment. This dependence involves a balance between analysis and intuition, where good judgment is the ability to go beyond the information given and rely on broader knowledge and experience [7]. If students have been receiving this experience, by virtue of participating in FTXs, then they should exhibit improved performance on the SJTs between pre-test and post-test.

The SJT instrument is a particularly useful tool because its realistic scenario items reasonably approximate the types of cognitive process improvements expected from repeated FTXs. By using the SME instructor ratings as the 'answer key', we assessed how the students' mental representations of various real-world problems began to resemble the instructors' mental representations.

For the Combat Tracking SJTs, our review identified six skill areas that were particularly important for success and amenable to testing via a short, written scenario: methods for closing the time/distance gap, executing lost spoor procedures, counter-tracking tactics, tactical formations, ground spoor characteristics, and dynamics of the footprint. For each area, a one-paragraph scenario was prepared to set up the problem. Then six alternative course of action options were generated that might be initiated by the tracker in response to the problem. In developing these courses of action, two options were specified that would be superior, two that would be inferior, and two that would have both strengths and weaknesses.

An example item from one of the Tracking SJTs is shown below. The scenario set-up is brief, reducing the reading requirement and leaving desirable information omitted. The student must make inferences and exercise judgment which is aided by the experience obtained during the FTX – the main intent of the SJT.

-----------Example SJT Item – Topic Area: Counter-tracking----------------

Your team has been tracking an experienced, well-armed band of insurgents for several days. The time/distance gap has been slowly closing to where it is now about 8 hours. You come upon where their tracks should be, but they have been obliterated by tracks of local cattle that cut through the ground spoor from several directions. Please rate the effectiveness of the following six decision options using this five-point scale. Don't hesitate to use the entire scale in judging these choices.

5 = highly effective/4 = moderately effective/3 = neutral/2 = moderately ineffective/1 = highly ineffective

[ ] Have one of your flanker trackers and the rear security tracker back track to the point where the cattle came from to see if the quarry's tracks are intermixed with them

[ ] Initiate a 360-degree lost spoor procedure

[ ] Look at surrounding tree branches in the immediate area for aerial spoor to estimate if/when the quarry had been there

[ ] Change to a Ranger/single file formation to look for any quarry ground spoor that might have escaped obliteration by the cattle

[ ] Change your tracking direction to follow the cattle path with the highest density of tracks

[ ] Slow pace of tracking movement to prepare for counter-tracking tactics

----

We used a similar procedure to create the Profiling SJT. In this case, we thoroughly reviewed our field notes from the Combat Hunter Limited Objective Evaluation, and identified 5 topic areas for focus: six combat domains (e.g., atmospherics, geographics), tactical cunning (think like the enemy), optical devices, tactical patience, and combat rules of 3. As with Tracking, we created two equally-difficult versions of the Profiling SJT, with six response options for each item that covered the range of effectiveness. Each SJT was administered to the students as a group during class time. The pre-test was given just prior to the first field scenario and the post-test was given on the day after the last field scenario.

For scoring, students' SJT answers were compared to those from the instructors who had also taken the SJT, which provided the answer key. To compute a score, we used the sum of the squared deviations from the instructors' response [8]. With this method, underestimates and overestimates of the SME rating are weighted equally, where extreme deviations are weighted more heavily. Higher scores mean worse performance since they are more discrepant from the SME's assessment. To determine degree of learning during the FTX, we computed each participant's difference score as their post-test score minus pre-test score; negative scores correspond to improved performance on the post-test.

3.2 Results

The SJT results are presented in Figure 1 for both class segments. The figure indicates the number of students whose post-test and pre-test difference score fell into one of the bins of size 10. Negative scores indicate a learning effect, as students' deviation score (from the instructors') was smaller on post-test (e.g., a desirable outcome).

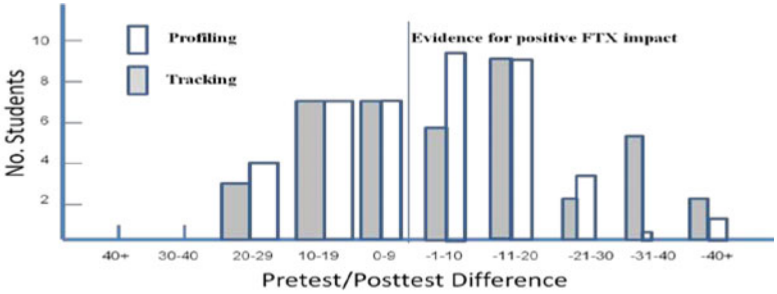


Fig. 1. Frequency distribution of SJT posttest/pretest difference scores

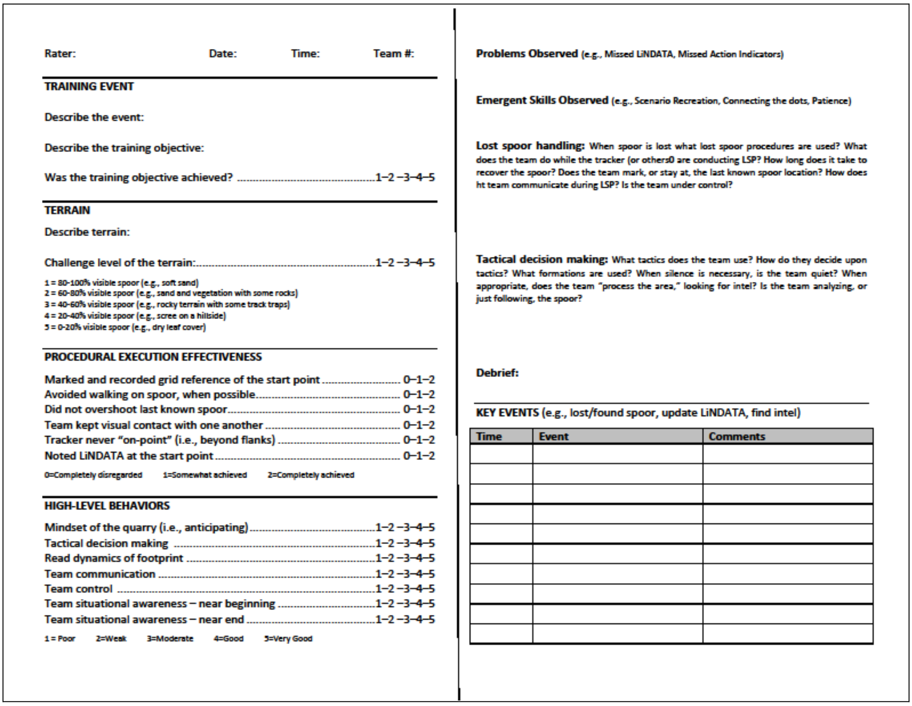
The frequency distribution gives us clues on the locus of the learning effect. For both course segments, there are no instances where students had high (30+) positive scores. On the right side of the Tracking distribution, seven participants had large decreases in their SJT score on the post-test, producing differences of -30 or greater. The FTX experience is to calibrate those students whose initial (pre-test) judgments were askew from the modal representation of the SME instructors. There was statistical evidence of learning as a paired t-test indicated that, on average, students' post-test scores were lower than their pre-test scores (t = 2.229, p < .011, df = 41).

The Profiling effects were similar though smaller in magnitude. This was most evident in the -31-40 bin, where Tracking had five students, but Profiling had 0. The Profiling t-test was not significant, though the trend was in the right direction (t = 1.114, p < .26, df = 39). Part of the reduced effect was likely due to test fatigue, as students took the SJT post-test on the last day of training [11]. On balance, the SJTs indicated that students acquired improved judgment and decision-making from FTX experience.

4 Behavioral Observations

4.1 Method

Behavioral Observation Checklists (BOCs) offer a structured method to collect quantitative and qualitative data on individual and team performance during FTXs. The BOCs used in BH were modified from ones used in the CH evaluation [3]. Separate BOCs were created for Tracking and Profiling, where the former is shown in Figure 2. The instrument was created in a two-column layout so it could be folded for portability. The upper left part of the BOC has space to describe the training event for that day. A set of 3-point rating scales gauge student proficiency in basic procedural skills whereas 5-point scales capture six high level behaviors. The right column lets the researcher comment on emerging skills, lost spoor handling (an important tracking skill), and decision-making. A similar format was used for the Profiling BOC.



Rater: Date: Time: Team #:

TRAINING EVENT

Describe the event:

Describe the training objective:

Was the training objective achieved? ....1–2 –3–4–5

TERRAIN

Describe terrain:

Challenge level of the terrain:..1–2 –3–4–5

1 = 80-100% visible spoor (e.g., soft sand)
2 = 60-80% visible spoor (e.g., sand and vegetation with some rocks)
3 = 40-60% visible spoor (e.g., rocky terrain with some track traps)
4 = 20-40% visible spoor (e.g., scree on a hillside)
5 = 0-20% visible spoor (e.g., dry leaf cover)

PROCEDURAL EXECUTION EFFECTIVENESS

Marked and recorded grid reference of the start point 0–1–2
Avoided walking on spoor, when possible.................................... 0–1–2
Did not overshoot last known spoor.. 0–1–2
Team kept visual contact with one another 0–1–2
Tracker never “on-point” (i.e., beyond flanks) 0–1–2
Noted LiNDATA at the start point 0–1–2

0=Completely disregarded 1=Somewhat achieved 2=Completely achieved

HIGH-LEVEL BEHAVIORS

Mindset of the quarry (i.e., anticipating)................................1–2 –3–4–5
Tactical decision making ....1–2 –3–4–5
Read dynamics of footprint ....1–2 –3–4–5
Team communication ....1–2 –3–4–5
Team control ....1–2 –3–4–5
Team situational awareness – near beginning ....1–2 –3–4–5
Team situational awareness – near end ....1–2 –3–4–5

1 = Poor 2=Weak 3=Moderate 4=Good 5=Very Good

Problems Observed (e.g., Missed LiNDATA, Missed Action Indicators)

Emergent Skills Observed (e.g., Scenario Recreation, Connecting the dots, Patience)

Lost spoor handling: When spoor is lost what lost spoor procedures are used? What does the team do while the tracker (or others0 are conducting LSP? How long does it take to recover the spoor? Does the team mark, or stay at, the last known spoor location? How does ht team communicate during LSP? Is the team under control?

Tactical decision making: What tactics does the team use? How do they decide upon tactics? What formations are used? When silence is necessary, is the team quiet? When appropriate, does the team “process the area,” looking for intel? Is the team analyzing, or just following, the spoor?

Debrief:

KEY EVENTS (e.g., lost/found spoor, update LiNDATA, find intel)

Time	Event	Comments

Fig. 2. BOC used to collect performance data during Combat Tracking

Performance data were collected by assigning one researcher to each team for each day of the Tracking FTXs. Researchers rotated teams over days to give ample exposure to the teams. A similar strategy was used for the Profiling FTXs. On each day, researchers met with the instructors to understand the objectives and goals of training for that day. In addition, two researchers were assigned to observe the same team on select days of the Tracking FTXs. This was done to allow an assessment of inter-rater reliability (Kappa = .59), which corresponded to ‘moderate agreement’.

4.2 Results

Pooling rating data across teams and days yields stable quantitative trends showing how student BPRA behaviors improved. To smooth out daily fluctuations, we pooled data from successive days or scenarios for the basic procedural skills and the higher-level behaviors. A typical result is depicted in Figure 3, where performance ratings for three high level Tracking behaviors (e.g., reading footprint dynamics, adopting a 'quarry mindset', and tactical decision-making) – are plotted across days. Statistical analysis revealed that, despite starting out at fairly high levels, all three measures showed significant increases ($p < .025$) from Days 2-3 to Day 10.

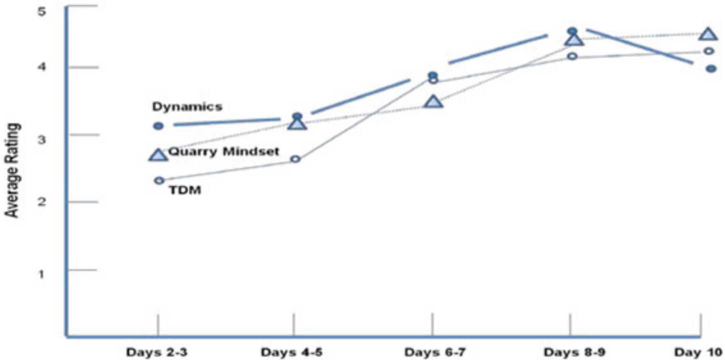


Fig. 3. Average performance ratings for three higher-level Tracking behaviors

Analysis of the quantitative data revealed that most procedural and higher-level behaviors exhibited similar increases across days or scenarios, indicative of a training effect. For Tracking, most of the basic procedural skills (3-point ratings) increased significantly across days, such as avoiding walking on the spoor line, maintaining visual contact, and marking the starting point of a track. For high level behaviors, besides those in Figure 3, we saw that students' performance on situation awareness, communication, and team control increased significantly ($p < .05$) over days. For Profiling, most procedural behaviors (e.g., spreading observations across team members, recording observations, using profiler language, establishing a stable baseline, using criteria to make a positive ID) improved over scenarios. All Profiling high level behaviors, rated on BOC 5-point scales, also improved. These included detecting basic events, adopting an insurgent mindset, interpreting complex events, communication, anticipating upcoming events, and exhibiting tactical patience.

The qualitative data on the BOCs, researcher comments, yielded key insights concerning the content of student behaviors, problems that emerged during the FTXs, and effective instructional techniques. Only a sampling of these results can be presented here; detailed findings are in [9]. For example, the Tracking BOCs show that early in training, students keep their heads mostly down so they can pick up the details of individual tracks (i.e., micro-tracking). While effective for seeing detail, it is slow. Later, students begin to look up more, using the pattern in the track line to discern where the quarry is likely headed (macro-tracking). This is a faster and more efficient technique. Many other aspects of tracking show similar qualitative trends.

Researcher comments on the Profiling BOCs also captured notable problems, trends, shifts in student behavior, effective instructional techniques, and emerging skills. Regarding the latter, students exhibited various skill improvements as they gained scenario experience, such as improved ability to identify high value individuals, synthesize events (i.e., 'connect the dots'), interpret complex events, and predict complex events from early signs. In later scenarios, even more complex skills were emerging, such as scenario recreation ability, trust building, anticipation to get even more ‘left of bang’, and adopting the mindset of other cultures.

5 Emerging KSA Profiles

The third behavioral category entailed applying a comprehensive framework of KSAs, developed previously for CH [3], to researcher field notes to capture the richness with which the students’ tracking and profiling competencies were emerging over the course of field training. Thirty-three KSAs were defined and organized into a six-category taxonomy [9]: use of enhanced observation techniques, identification of critical event indicators, interpretation of human behavioral cues, synthesis of ambiguous information, proactive analysis and dynamic decision-making, and employment of cognitive discipline. These KSAs apply equally to Tracking and Profiling, as they are both focused on a common set of IW cognitive and behavioral processes; only their manifestations as behavioral markers differ. An example KSA from Cognitive Discipline is shown below.

Table 1. Example of a Knowledge-Skill-Attitude

KSA	Behavior Marker – Profiling	Behavior Marker - Tracking
22. Keep an open mind to the unexpected (recognize there are unknown variables in the situation)	Do they consider the possibility that insurgents might use new tactics (e.g., different IED emplacing) or attempt something completely different than anything that has been tried before?	Do they consider that the hostiles might consider something completely different, like splitting up to rejoin at a rally point further down the track line?

The KSAs served several purposes in BH. They provide a framework to represent course content so training objectives from loosely-defined field scenarios could be established. KSAs link with training outcomes, and are observable, measurable, and trainable. They link Profiling and Tracking, establish validity of the training effort, and facilitate packaging of course materials for other venues.

Using this framework, three researchers took their field notes and instantiated all 33 KSAs for both course segments. A complete, emergent profile was developed that indicated the team (1-5), behavior observed, and day when the behavior occurred. Inter-rater agreement was high and the framework applied to both course segments, only the specific behaviors were different. These profiles were then employed by the larger research team to create a comprehensive program of instruction for BH [4].

6 Conclusions

Collectively, the three classes of behavioral measures reported here – SJTs, BOCs, and KSAs – form a powerful methodology to capture student performance during

FTXs. During BH training, BPRA competencies were assessed using the controlled testing conditions of SJTs, repeated quantitative elements of BOCs, and comprehensive qualitative KSA profiles. Applicable to both Profiling and Tracking, these behavioral measures are well-suited for "reading" the human terrain in CT operations. We believe that all three behavioral methods will be highly useful in other applications where field training of CT skills is taking place. When utilized by well-trained, experienced researchers, and with an appropriate level of setup and incorporation of subject matter expertise into the materials, SJTs, BOCs, and KSAs offer a highly effective method to assess training performance, capture training objectives, and identify the most successful instructional approaches for counter terrorism.

References

1. Department of Defense Directive 3000.07: Irregular Warfare (December, 2008), `http://www.dtic.mil/whs/directives/corres/pdf/300007p.pdf` (Retrieved from June 28, 2010)
2. Kobus, D., Williams, G.: Training tactical decision making under stress in cross- cultural environments. In: Proceedings of the Conference on Cross-Cultural Decision Making, Miami, FL [CD-ROM] (2010)
3. Spiker, V.A., Johnston, J.H.: Limited objective evaluation of combat profiling training for small units. (Technical Report). US Joint Forces Command, Suffolk, VA (January 2010)
4. Institute for Training and Simulation. In: Schatz, S., Fautua, D.(eds.): Border Hunter research technical report. US Joint Forces Command, Suffolk, VA (July 2010)
5. Gessner, T.L., Klimoski, R.J.: Making sense of situations. In: Weekley, J.A., Ployhart, R.E. (eds.) Situational judgment tests. LEA, Mahwah (2006)
6. McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., Braverman, E.P.: Use of situational judgment tests to predict job performance: A clarification of the literature. Journal of Applied Psychology 86(4), 730–740 (2006)
7. Brooks, M.E., Highhouse, S.: Can good judgment be measured? In: Weekley, J.A., Ployhart, R.E. (eds.) Situational judgment tests. LEA, Mahwah (2006)
8. Weekley, J.A., Ployhart, R.E., Holtz, B.C.: On the development of situational judgment tests: Issues in item development, scaling, and scoring. In: Weekley, J.A., Ployhart, R.E. (eds.) Situational judgment tests. LEA, Mahwah (2006)
9. Spiker, V.A., Johnston, J.H.: Border Hunter: Evaluation of field training (Technical Report). US Joint Forces Command, Suffolk, VA (July 2010)

Part VI
Aerospace and Military Applications

Applied Cognitive Ergonomics Design Principles for Fighter Aircraft

Jens Alfredson[1], Johan Holmberg[1], Rikard Andersson[1], and Maria Wikforss[2]

[1] Saab AB, Aeronautics, SE-581 88 Linköping, Sweden
[2] Saab AB, Security and Defence Solutions, SE-175 88 Järfälla, Sweden
{Jens.Alfredson,Johan.Holmberg,Rikard.Andersson,
Maria.Wikforss}@Saabgroup.com

Abstract. The objective of the reported work was to study the use and applicability of applied cognitive ergonomics design principles for fighter aircraft, with examples from the modern Swedish swing-role aircraft Gripen. Methods used were a literature review of relevant design principles together with an analysis of their applicability to the fighter aircraft domain as well as interviews of developers and scrutinized system documentation of ongoing fighter aircraft development at Saab. As a result of those activities, we can here present a brief description of cognitive ergonomics design principles applied in the Gripen fighter aircraft, and the development process for human-machine interaction for fighter aircraft. Finally, considerations for the design process for fighter aircraft are discussed in the context of that description.

Keywords: Fighter Aircraft, Design Principles, Cognitive Ergonomics, Human-Machine Interaction.

1 Introduction

For fighter aircraft it is very important to achieve good human system integration. It is important to regard human factors in, for instance, display design, for the fighter pilot and the aircraft to perform optimal together. This is important for safety reasons, since flying an aircraft is in itself an activity that is potentially dangerous and even more so in a hostile environment. Also, it is important to regard human factors for reasons of performance which are central for fighter aircraft. The pilot interaction is often essential and is sometimes a bottleneck in the decision making process involving decision support systems and other automation with human-in-the-loop decisions under high demands, such as high mental workload and situations requiring high situational awareness.

The cockpit design, including display design and interaction design has evolved through a process including various amounts of elements such as, design tradition, user involvement, structured design processes, human factors knowledge and more. As part of a local initiative for efficient development of fighter aircraft, the reported study was performed, focusing on efficient development of fighter aircraft interaction

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 473–483, 2011.

design, including methods and examples of usability design principles in the modern Swedish swing-role aircraft Gripen [1].

The objective of the reported work was to study the use and applicability of applied cognitive ergonomics design principles for fighter aircraft, with examples from the Gripen aircraft. By the use of methods for literature review, interviews, and analysis, descriptions are presented of cognitive ergonomics design principles applied in the Gripen fighter aircraft, and the development process for human-machine interaction (HMI) for fighter aircraft, below.

This work is not only relevant for the domain of fighter aircraft development, but also for other application areas where design principles within cognitive ergonomics are to be adopted into applications where successful human system interaction is central. However, the fighter domain has specific constraints to be regarded when studying the applicability of cognitive ergonomics design principles.

2 Fighter Domain Constraints

Even though the demands of the fighter aircraft domain are very special, just as many other domains are special in their own sense, general design principles were found to be useful for describing the pilot-aircraft interaction in this case, which supports the idea of general design principles, although it is important to be aware of the specific requirements of the domain. By designing for what is special about a fighter pilot, a fighter aircraft and the flight environment relevant design concepts can be formed from design criteria [2].

Fighter pilots are a homogeneous group, especially compared to users of consumer products, e.g. mobile phones or cars. The group contains neither old users, nor very young, and they are selected thoroughly with regards to, for example, mental capabilities and anthropometrical characteristics. They have all successfully carried out the same demanding pilot training and thereby acquiring the skills and procedures needed. While it is, of course, easier to design for a homogenous user group it is important to stress the fact that the demanding context for a fighter pilot still provides extreme challenges from a design point of view.

There are high demands on a fighter pilot regarding rapid decision-making. Decisions must be made under extreme time pressure in a hyper-dynamic setting in a hostile environment, typical for naturalistic decision making [3]. There are several aspects that are important for a fighter aircraft domain, and many of them could be said to be extreme. Some examples are:

- High G-loads
- High mental workload
- Sudden and drastic light conditions and high visual demands
- Demands on rapid decision-making in a battle of life and death

In scenarios with high G-loads the pilot's interaction possibilities are degraded which needs to be considered during the design process.

3 Method

Methods used were a literature review of relevant design principles together with an analysis of their applicability to the fighter aircraft domain as well as interviews of developers and scrutinized system documentation of ongoing fighter aircraft development at Saab. A literature review on usability design principles and interviews of system developers at Saab [1] were complemented with further analysis to find suitable descriptions of the development process for HMI for fighter aircraft. Below the methods used are described, and the outcome is presented in following chapters.

3.1 Interviews

Interviews were initially performed with four very senior developers of HMI for fighter aircraft at Saab. They were all men between 42 to 65 years of age, and each had between 17 to 40 years of experience of the domain. First, all respondents were interviewed through topic focused, semi-structured interviews at one or two occasions each. To ensure that the respondents were to address the same topics, themes for the interview sessions had been prepared in advance. The notes from each interview were fused with the other notes and were validated and further elaborated in a group interview/workshop with all but one of the respondents present. These interviews were specifically concerned with design principles, and were followed by unstructured interviews to complement the analysis of the development process.

3.2 Literature Review

A review of the research field of relevant design principles was conducted. Also, system documentation of ongoing fighter aircraft development at Saab was scrutinized.

3.3 Analysis of Design Principles and Development Process

Based on the interviews and the literature review an analysis of design principles was performed. Design principles from literature were systematically mapped towards the outcome of the interviews and described in examples of the type presented below. Similarly, but less structured, an analysis of the development process were performed based on the outcome of unstructured interviews.

4 Cognitive Ergonomics Design Principles Applied in the Gripen Aircraft

The examples below are in part based on analysis of usability design principles for the Gripen aircraft [1], but are also selected to be suitable examples for a reader that does not necessarily have experiences from the domain. One category of design guidelines found in the literature and in the interviews had to do with *Consistency* (Fig. 1, Fig. 2, Fig. 3, and Fig. 4 show examples). Examples of these guidelines were: what look alike, should act alike [4], the same actions should lead to the same result [5], similar situations should be handled similarly [6], be consistent in automation [7], and be consistent in formatting, terminology, positioning, attention grabbers, etc. [8].

Another category is related to *Support of user mental models* (Fig. 2, Fig. 3, and Fig. 8 show examples). Examples of these guidelines were: use clearly defined conceptual models in display design [9], mirror the user's mental model - not the designer's, mimic well known concept and such that is previously learnt, talk the user's language, give feedback, avoid irrelevant information, develop for both experts and novices but protect the novices from complexion [5], all the information needed for a task should be available when performing the task [10], use system visibility, and what moves on the screen should follow the user's mental model of what actually moves [9].

A third category had to do with *Keep it simple* (Fig. 2, Fig. 5, Fig. 6, and Fig. 7 show examples). Examples of these guidelines were: eliminate what doesn't add to efficient use [4, 5], information used sometimes, shouldn't be displayed always [11], use visual hierarchy [4, 5], use default settings, more common tasks should be made easier [5], display data in a directly usable format for the user [12], limit the number of separation lines [13], simplify symbols as far as possible while keeping the understanding [14], group primary information on one display [15], minimize user options [5], and group information that is integrated mentally together [9].

A fourth category is related to Use of color (Fig. 8 shows an example). Examples of these guidelines were: design for monochrome, add color only as redundant information [4, 5, 6, 15, 16], limit number of colors to about 5±2 [4, 6, 14, 15, 16], avoid overuse of color (noise) cf. [17], follow user expectations and domain color usage [5, 6] use conventional colors for danger/warning/normal – red/yellow/green [17], be consistent in color usage throughout the system [4, 16], for bright background use low-intensity colors such as off-white, for dark background use cool colors like black or blue, foreground and background should have good contrast [5], use sharp colors to grab attention [16].

More categories like "Multimodality/redundancy" were also found in literature and application but will not be described here.

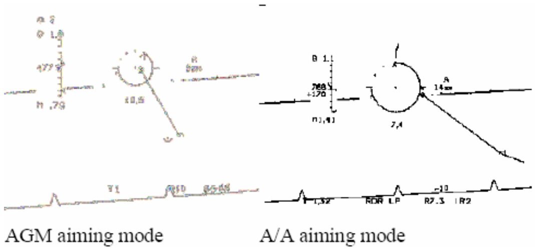


Fig. 1. HUD Aiming modes (consistency)

With the introduction of Gripen as a multirole fighter in the Swedish air force the same pilot was supposed to perform several types of missions previously done by specialized pilots. In order to ease the training the HMI for similar tasks were designed to be as similar as possible rather than being optimized for each task separately. Examples of this are the gun and missile aiming modes, where similar display and controls are used regardless of target types like aircrafts, ships, buildings or land vehicles.

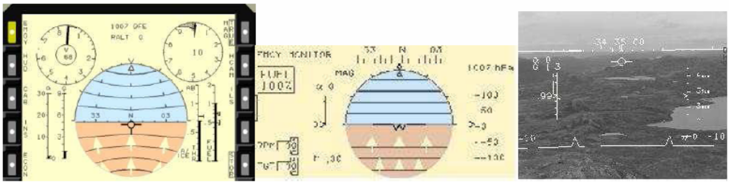


Fig. 2. Basic T (Consistency, Support of user mental models, Keep it simple)

Consistency in design because of pilot training is nothing new. During World War II and in the early days of the cold war new aircraft types were being introduced at a remarkable rate. The pilots then had to convert from one aircraft type to another and then relearn the user interface. During stress humans have a tendency to revert to the basics they learnt and when the user interface differed accidents could happen. In order to mitigate this, the "Basic-T" concept was created as a design guideline for how the primary flight data instruments should be located relative each other in the cockpit. Even modern day electronic flight instruments adher to the "Basic-T" concept.

The mental model for attitude display in Gripen is to think of the aircraft as positioned inside a giant fixed sphere where each ten degrees of pitch and heading is indicated. It is also an example of the "inside out" conceptual model used throughout the Gripen displays.

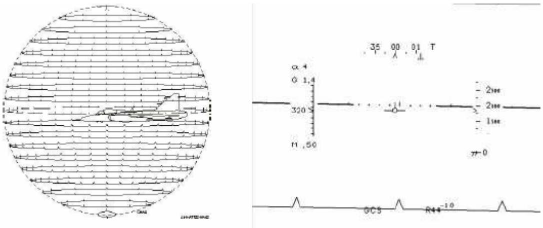


Fig. 3a. Attitude display (Consistency, Support of user mental models)

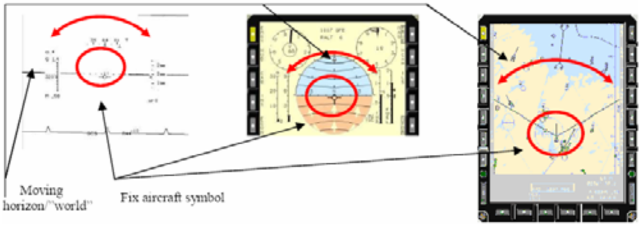


Fig. 3b. Attitude display (Consistency, Support of user mental models)

The design guideline “what moves on the screen should follow the user’s mental model of what actually moves” is one of the reasons behind the “inside out” conceptual model and as can be seen affects not only the Head Up Display but map displays and Attitude Direction Indicators.

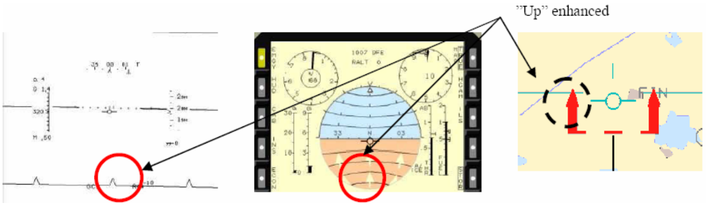


Fig. 3c. Attitude display (Consistency, Support of user mental models)

The way “up” is enhanced for potentially dangerous attitudes is both consistent and complies with the mental model of attitude.

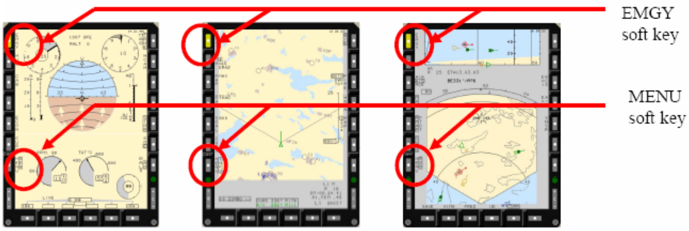


Fig. 4. Soft key positioning (Consistency)

Standard soft keys are located at the same position on all displays and have same or similar results when activated.

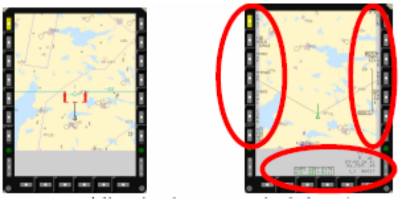


Fig. 5. Hiding irrelevant tactical data (Keep it simple)

When ground collision warning is given the pilot’s focus should be on saving the aircraft so all tactical data is turned off on the more information intense displays.

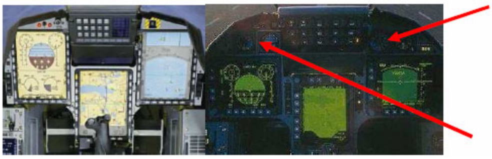
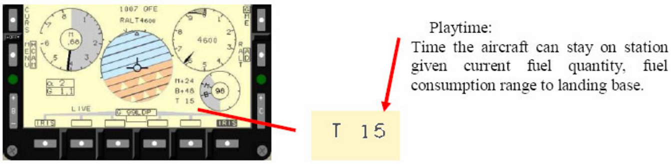


Fig. 6. Emergency instruments (Keep it simple)

In Gripen C/D there are no longer any dedicated standby instrument, instead every display can act as an emergency flight data display directly connected to the sensors.



Playtime:
Time the aircraft can stay on station given current fuel quantity, fuel consumption range to landing base.

Fig. 7. Playtime (keep it simple)

An example of displaying data in a directly usable format for the user is the display of “playtime” rather than displaying fuel quantity and fuel consumption rate that was the norm in older aircraft.

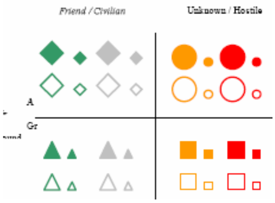


Fig. 8. Target symbols (Color, Support of user mental models)

The target symbols are designed to work even when color is hard to see when display lighting is turned low at night to preserve the pilot’s night vision. Color is however used to enhance the target symbols, making them easier to group. A target symbol can represent data from many different sensors but these are fused into a single display symbol to support the pilot’s mental model of “one object”.

5 Development Process for Human-Machine Interaction for Fighter Aircraft

Positive qualities of formalization of applied cognitive ergonomics design principles were found for the integration of human factors considerations into the development process of fighter aircraft. However, the development process was found to be only partially influenced by applied cognitive design principles and we could not fully describe the impact of other influences of the design, in the design process at hand. However, the use of style guides were found to be used as a means of implementing design principles into the design process.

The design process for HMI is a central part of the development process for fighter aircraft. The design process itself can be designed, through both technological design as well as institutional design [18], since in parallel to an architect that designs a building, an organization could also be designed [19]. It is important to regard user needs when the design process is formed, either by direct involvement by users or indirectly by for instance descriptions of user behavior or general design principles. Often a designer regards the user as subsystem of the total system and is perhaps modeled through design principles or scenario elements, depending on the types of meetings in the applied design process [20].

The development process at Saab for HMI has for a long time been concerned with simulator-based design in the development of Swedish fighter aircraft [21]. Saab has developed a process with emphasis on rapid prototyping and found it to be successful. When reassessing the characteristics of the current design process simulator-based design was still found to be central. Early and frequent user involvement in an iterative design process are still heavily influencing the design, perhaps more than formal descriptions of cognitive ergonomics design principles. Singer [22] has presented how part-task simulation in an early part of the design process was useful for designing commercial aircraft cockpit systems. In the development process for HMI for fighter aircraft there is a special role for a style guide to apply cognitive ergonomics design principles.

Since HMI design work is performed in close connection to various functions in cockpit there is a risk that the design will diverge between different design teams which leads to a total design that is inconsistent. One tool used to mitigate that risk is a HMI style guide. The HMI style guide is used to document and distribute general design decisions and it contains the design philosophies and design rules that are the foundation of work within the HMI design in the cockpit. Examples regarding the design philosophies are support of mental models, redundancies and manageable work load. The design rules are to be followed when working with the development of functions that involves interaction between user, usually the pilot, and the system. The content of the style guide should be based on relevant ISO and Military standards, public HMI guidelines and own experiences from development of aviation systems. Updates are conducted regularly based on experience made in ongoing development and changes or supplements in standards and international HMI research.

One risk when inserting a style guide is that it will be seen as a guarantee for a usable system and that its existence makes other essential design activities unnecessary. It is important to stress the fact that the style guide shall be regarded, and managed, as a support to the HMI-developers and not as a template of making a useful system and, accordingly, does not lessen the need for a user centered design process with its associated activities.

6 Discussion and Conclusions

This study has shown, on the level of provided examples, that cognitive ergonomics design principles are applied for fighter aircraft. Further, the applicability of cognitive ergonomics design principles in the design process was studied. It was found that cognitive ergonomics design principles influence the design, directly or by the use of descriptions such as style guides. However, it was also found that, in the studied development context, the design principles were processed and formed in the design process, so that a specific design principle was unlikely to be uncritically implemented, automatically, but rather by a compromise with other input to the design process.

Even though the focus of this study was cognitive ergonomics design principles it may be interesting to reflect on what this complementary input to the design process is. For instance, standards influence the design, design tradition influence the design, as well as contextual demands and technical or economical constraints on the design process. Also, there may be conflicts between different design principles for a specific design decision. For instance, the principle of consistency often is in conflict with other design principles describing their “local good” (optimizing for that specific design principle).

An example of when the principle of consistency could be in conflict with design tradition is that most often imperial units are used in aircraft and metrical units are often used by ground based army units. However, when they are cooperating, for instance in a close air support mission, the designer of the pilot interface has to handle that.

Contextual constraints for the fighter domain, as described earlier, are central for the design of pilot interfaces. An example of a conflict between the contextual constraints and a design principle could be when designing for access of two types of sensors. If one sensor has to have high access, since it, for instance, is critical for dogfight it may not be designed to be handled in the same way as another senor with low access demands, used for, for instance reconnaissance. Since standards partly are based on design principles you may think that there would never be a conflict between what is stated in a standard and a design principle. However, this is not always the case. For instance, if a civil standard for radio navigation is using a lot of colors it is hard to both follow that standard and at the same time optimize the use of colors for the rest of the interface, and at the same time keep consistency throughout the design.

The perhaps most common conflict is the one between a design principle and technical and economical constraints. Especially technical limitations are often hard to change even if they are in contrast with a design principle. For instance, at the time when no color displays had been developed and tested for fighter aircraft principles for use of color could be in conflict with those constraints.

However, in this study not only conflicts were found between fighter domain constraints and cognitive design principles applied in the Gripen aircraft, in the development process for HMI for fighter aircraft. Positive qualities of formalization of applied cognitive ergonomics design principles were found for the integration of human factors considerations into the development process of fighter aircraft. The performed study has influenced the HMI design process at Saab. For instance, the use of style guides has been improved, as one tool among others to link general design principles to the demands of a specific design process.

To conclude, this work has studied the relevance of applied cognitive ergonomics design principles for fighter aircraft. Design principles are relevant for human factors considerations in development, although design principles are only one of many sources of information for a designer. For instance, the domain specific design tradition regards what is unique for the domain, so the general design principles has to be adjusted to what is unique for the domain to be competitive in domain specific human factors evaluations, such as human-in-the-loop simulations.

Future research is needed to develop methods and approaches for enhanced use of applied cognitive ergonomics design principles for fighter aircraft. Recently a national research project has been started supported by The Swedish Governmental Agency for Innovation Systems (VINNOVA) through the National Aviation Engineering Research Program focusing on innovative and effective system development for military aircraft studying how qualitative and quantitative assessment of pilot decisions can be applied through formal descriptions of design in a so called "Brainbudget", such as time to decision, amount of mental workload etc. Future research could also be focused on comparisons between various development contexts, such as other domains, to better understand how the findings from this work compares to related work with other constraints.

References

1. Wikforss, M.: Usability Design Principles in JAS 39 Gripen. Master Thesis (No. 2008:121). Royal Institute of Technology, Stockholm, Sweden (2008) (in Swedish)
2. Alfredson, J., Andersson, R.: Managing Human Factors in the Development of Fighter Aircraft. In: Abu-Taieh, E., El Sheikh, A., Jafari, M. (eds.) Technology Engineering and Management in Aviation – Advancements and Discoveries. IGI Global (in press)
3. Klein, G., Orasanu, J., Calderwood, R., Zsambok, C.: Decision Making in Action – Models and Methods. Ablex, Norwood (1993)
4. Marcus, A.: Graphical User Interfaces. In: Helander, M.G., Landauer, T.K., Prabhu, P.V. (eds.) Handbook of Human-Computer Interaction - Second, Completely Revised Edition, pp. 423–440. Elsevier Science B.V., Amsterdam (1997)
5. Galitz, W.O.: The Essential Guide to User Interface Design - An Introduction to GUI Design Principles and Techniques, 3rd edn. Wiley Publishing, Indinapolis (2007)
6. Shneiderman, B., Plaisant, C.: Designing the user Interface - Strategies for effective Human-Computer Interaction, 4th edn. Addison-Wesley, Boston (2005)
7. Endsley, M.R., Bolté, B., Jones, D.G.: Designing for Situation Awareness - An Approach to User-Centered Design. Taylor & Francis, New York (2003)
8. Faulkner, C.: The Essence of Human-Computer Interaction. Pearson Education Prentice Hall, Edinburgh Gate, England (1998)
9. Wickens, C.D., Lee, J.D., Liu, Y., Gordon Becker, S.E.: An Introduction to Human Factors Engineering, 2nd edn. Pearson Education Prentice Hall, Upper Saddle River (2004)
10. Bainbridge, L.: Multiplexed VDT Display Systems - A Framework for Good Practice. In: Weir, G.R.S., Alty, J. (eds.) Human-Computer Interaction and Complex Systems, pp. 189–221. Academic Press, San Diego (1991)

11. Preece, J., Rogers, Y., Sharp, H.: Interaction design – Beyond Human-Computer Interaction. John Wiley & Sons, New York (2002)
12. Williges, R.C., Williges, B.H., Fainter, R.G.: Software Interfaces for Aviation Systems. In: Weiner, E.L., Nagel, D.C. (eds.) Human Factors in Aviation, 2nd edn., pp. 463–494. Academic Press, San Diego (1988)
13. Löwgren, J.: Human-Computer Interaction - What Every System Developer Should Know. Studentlitteratur, Lund, Sweden (1993)
14. Sanders, M.K., McCormick, E.J.: Human Factors in Engineering and Design, 7th edn. McGraw-Hill, Singapore (1993)
15. Nielsen, J.: Usability Engineering. Academic Press, London (1993)
16. Andrén, P., Gunnarsson, S., Lundin, J.: Grafiska användargränssnitt - en utvecklingshandbok. Studentlitteratur, Lund, Sweden (1993) (in Swedish)
17. Wagner, E.: System Interface Design - A Broader Perspective. Studentlitteratur, Lund, Sweden (1994)
18. Koppenjan, J.F.K., Groenewegen, J.P.M.: Institutional Design for Complex Technological Systems. Int. Journal of Technology, Policy and Management 5(3), 240–257 (2005)
19. Simons, R.: Levers of Organization Design – How Managers Use Accountability Systems for Greater Performance and Commitment. Harvard Business School Press, Boston (2005)
20. Darses, F., Wolff, M.: How do designers represent to themselves the users' needs? Applied Ergonomics 37(6), 757–764 (2006)
21. Alm, T.: Simulator-Based Design – Methodology and Vehicle Display Applications. Doctoral Dissertation (No. 1078), Linköping University, Linköping, Sweden (2007)
22. Singer, G.: Methods for Validating Cockpit Design – The Best Tool for the Task. Doctoral Dissertation. Royal Institute of Technology, Stockholm, Sweden (2002)

Designing Effective Soldier-Robot Teams in Complex Environments: Training, Interfaces, and Individual Differences

Michael J. Barnes[1], Jessie Y.C. Chen[1], Florian Jentsch[2], and Elizabeth S. Redden[1]

[1] U.S. Army Research Laboratory – Human Research & Engineering Directorate
Bldg 459, Aberdeen Proving Ground, MD 21005, USA
[2] University of Central Florida – Department of Psychology
Orlando, FL 32816, USA
{michael.j.barnes,jessie.chen,elizabeth.redden}@us.army.mil,
fjentsch@mail.ucf.edu

Abstract. Extensive US Army programs are being pursued to increase the effectiveness of unmanned vehicles for diverse missions during future combat. The following paper identified 23 human-robot interaction (HRI) guidelines related to interface design, procedural issues, individual differences and training implications based on three HRI research programs. The programs range from simulation experiments that investigated robot control in a multitasking environment from a mounted combat vehicle, to reconnaissance missions in a miniature Iraqi city that focused on Soldier–robot teaming relationships, to field studies at Ft. Benning that examined interface design issues for Soldiers supervising or controlling small robots.

Keywords: HRI design, military, human factors.

1 Introduction

The U.S. Army is engaged in an extensive research program whose goal is to develop unmanned vehicles (UVs) including both ground (UGV) and aerial versions (UAS) to act as force multipliers, to increase tactical flexibility, and most important, to save Soldiers' lives. While the eventual goal is autonomy, the current and near-term systems require teleoperations or waypoint control making the Soldier responsible for controlling the UVs and adapting to change. But even more important than control is the Soldier's role in understanding the mission environment both in terms of specific goals and meta-goals. For example, Robin Murphy [1] in reviewing real-world experiences with robotic systems points out that situation awareness (SA) rather than teleoperations was the most difficult problem for operators to overcome in finding survivors during the World Trade Center disaster. The purpose of this paper is to review three research programs funded by the Army in order to derive general principles and guidelines. These three programs were chosen because of their diverse approaches as well as their extensive findings over six or more years of research. All the programs were realistic in the sense that they dealt with real Army problems in

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 484–493, 2011.

operational settings. For each effort, we evaluated the implications of their results and derived general guidelines/ principles for designing UV systems that enhance Soldier HRI performance in diverse mission environments. We briefly describe the three paradigms, their most important results as well as guidelines for interface design, Soldier procedures, required skill sets, and training implications.

2 Simulation Studies at ARL-Orlando

The simulations conducted in Orlando were designed to investigate operator control or supervision of robots from a mounted vehicle during multitasking mission segments. The first four experiments emphasized difficulties mounted crews encountered when performing their normal duties in addition to remote targeting with a robot. The simulation studies manipulated both multi-tasking variables and degree of robotic control using a OneSAF simulation of a maneuver scenario in which the operator had to detect targets near the manned vehicles (primary task) and remote targets viewed from the robot's video (secondary task) as well as a communications and SA tasks. The initial experiment examined span of control issues whereas the latter three experiments manipulated variables related to the gunner conducting the robotic and gunner functions using separate displays. More important from a theoretical point of view was the impact of individual differences related to perceived attentional control (PAC) and spatial ability (SpA). Two of the experiments manipulated mitigation factors such as tactile cueing, aided target recognition (AiTR), and reliability - including both miss prone and false alarm prone AiTRs.

Three later experiments investigated the use of an intelligent agent, RoboLeader, to control 4-8 robots allowing the operator to supervise the agent instead of being required to control individual robots. The RoboLeader experiments were simulated in the MIX Testbed with an emulated operator control unit showing both map and streaming video views from the individual robots. Variables manipulated included number of robots, agent vs. direct control, individual differences, workload, SA, target density, and reliability (miss- and false alarm- prone).

2.1 Summary of Results

The initial mounted study indicated that operators were not efficient in using more than a single asset concurrently. They tended to over-rely on UASs compared to semi-autonomous UGVs although performance was approximately the same under single asset conditions [2]. In the gunner's study, operators performed more poorly on both the primary local security task and the secondary robotic remote targeting task as a function of robotic task difficulty [3]. In the third experiment, AiTR improved targeting on the primary task but had mixed results on the robotic task. Contrary to previous research, participants had better robotics task performance when they teleoperated the robot than when the robot was semi autonomous - suggesting over-reliance on automation for robotic tasks although the participants were aware of the limitations of the AiTR system on the robots [4]. In general, the experiments showed pronounced individual differences with participants with low SpA abilities and low PAC scores performing more poorly than those with better scores. However, there were some interesting interactions; the AiTR was more beneficial to the low SpA participants raising targeting accuracy nearly to those levels of the more proficient SpA participants; there was a similar trend for low PAC participants [4]. Also, low SpA participants tended to prefer visual cueing whereas high SpA participants were satisfied with tactile cueing. The PAC interactions were even more pronounced in the AiTR reliability study [5]; high PACs tended to perform poorly when using false-alarm-prone aids whereas low PACs tended to perform more poorly on the miss-prone aids showing a classical type X interaction indicating possible mistrust in the former case and over trust in the latter case. The RoboLeader studies investigated the usefulness of an intelligent agent to help supervise multiple robots [6][7]. RoboLeader supervision vs. operator supervision did not result in improved target detection for both the 4 and 8 robot conditions but RoboLeader supervision did result in more rapid transit for the robots. [6]. In the second intelligent agent experiment, the type of unreliability of the RoboLeader's route change suggestions had unexpected results [7]. The false-alarm-prone conditions (suggesting new routes when it was not necessary) actually resulted in better performance than the conditions with miss-prone suggestions (RoboLeader failed to suggest a new route when it was appropriate) because the false-alarm-prone suggestions could be easily checked by viewing the map but missing suggestions were more difficult to verify. Again, individual differences were important; high SpA participants performed better in scanning and target detection whereas high PAC participants' showed superior performance in the secondary tasks – communications and gauge monitoring [7].

2.2 Guidelines

1. Teleoperation creates significantly more workload for the operators than does other robotics tasks. In multitasking environments, operators may overlook other concurrent tasks or overly rely on automation for other tasks when teleoperating a robot. Although many of the ground robotic assets in the future will be semi-autonomous, teleoperation will still be an important part of any missions involving robotic assets (e.g., when robots encounter obstacles or other problems). The higher workload associated with teleoperation needs to be taken into account when designing the user interfaces for the robotic assets (see [8] for potential user interface designs).
2. Unreliable automated systems tend to affect operator's performance of concurrent tasks involving visual scanning and his/her situation awareness of the overall tasking environment [7] due to the operator's having to constantly monitor the automated system. Cueing displays in conjunction with aided target recognition capabilities should be implemented to enhance operator's overall task performance.
3. Participants' spatial ability (SpA) was found to be a reliable predictor of their targeting task performance. When selecting personnel for operating robotics systems for reconnaissance tasks, operators' SpA should be one of the criteria under consideration. This is consistent with conclusions of recent U.S. Air Force studies on the required abilities of UAS pilots and sensor operators [9].
4. The data from the studies conducted by Chen and her colleagues showed that those with lower SpA tend to prefer visual cueing over tactile cueing (in a visually intensive, multitasking environment), and those with higher SpA tend to favor

tactile cueing over visual cueing. To better enhance the task performance for low SpA individuals, the visual cueing display should be more integrated with the visual scene. For example, augmented reality (i.e., visual overlays) is a potential technique to embed directional information onto the video.

5. Operators' attentional control ability appears to have an impact on their reliance on automation and responses to unreliable alerts, especially when workload is heavy (e.g., teleoperating a robot while multitasking). Lee and See [10] suggested that training can be developed to educate the operators regarding the alert's "expected reliability, the mechanisms governing its behavior, and its intended use" (p. 74).
6. When selecting personnel for positions with frequent multitasking demands, attentional control abilities should be considered. A recent U.S. Air Force study concluded that attention control is one of the critical abilities of UAS pilots [9]. Consistent with this finding, the results in communication task performance in the studies conducted by Chen and her colleagues showed that those with better attentional control consistently outperformed those with lower attentional control. If personnel selection is not feasible, then training programs should be developed to improve the attention management skills of the low PAC operators to support multitasking requirements.

3 Human-Robot Teaming Research at Univ. of Central Florida

The research conducted at University of Central Florida (UCF) focused on the performance of teams of multiple operators as they interacted with one or more robotic vehicles. In all studies, of which there were over 12 between 2004 and the end of 2010, the teams' task was to perform Reconnaissance, Surveillance, and Target Acquisition (RSTA) in a mixed urban/suburban environment, represented by a custom-built 1:35 toy city prototypical of urban environments in Iraq [11]. The robotic assets in the simulation were UGVs that in size and mission resembled the Army's planned Armed Robotic Vehicle (ARV) (i.e., a vehicle of roughly the size of a small truck or small armored fighting vehicle). They were operated by confederate experimenters in a "man-behind-the-curtain" fashion; that is, the vehicles were operated by the confederates according to a script that specified the level of autonomy (LOA) while the research participants believed that the vehicles were real robots. Additionally, several studies employed UASs, which were simulated by planar and oblique camera(s) moving above the toy village in patterns similar to reconnaissance patterns employed by real UASs.

The conceptual frameworks underlying the research were (a) the Team Effectiveness Model (TEM), an input-process-output model of team processes, and (b) a conceptualization that put team performance of the mixed human-robot team at the top of a hierarchy with three underlying, or enabling, topics necessary to achieve team performance. The three underlying topics were: (a) Mutual planning and shared mental models in teams; (b) Self-localization and SA while using robotic/unmanned vehicles; and (c) Target recognition, friend-foe discrimination, and vehicle identification/classification. Only when the teams were able to correctly and efficiently perform all three underlying tasks, we posited, would they be able to engage in successful team performance. Different tracks within the research program, therefore, investigated the underlying topics and their determinants.

3.1 Summary of Results

Although the number of studies precludes a discussion of all results obtained through this research program, a number of highlights stood out in each area. First, conducting RSTA missions with UVs is an inherently difficult task; even well-trained teams with a military background found it both effortful and difficult to monitor the vehicles' paths, maintain spatial and situation awareness, detect and identify targets, and report them, while monitoring system health, reacting to unforeseen situations, etc. Further, precisely because the RSTA task is an attention- and memory-intensive task, we found that the addition of team members was helpful and generally improved performance, regardless of whether we changed from individual operators to teams of two, or from a team of two operators to a recon team with two operators and one mission manager. Indeed, under difficult circumstances and high workload, the addition of a team member improved performance by more than 1 person unit. In one study, the addition of a team member to an individual operator, thereby creating a team of two operators, improved performance by 190%. Clearly, the RSTA task is, at least at the current LOA, resource-intensive enough that the gains in performance from the addition of team members far outweigh the process losses typically associated with the communication and coordination necessary for team performance.

Our studies, however, did provide a more detailed picture than the mere statement that "adding team members is better." These studies also showed that technical, procedural, and individual factors all impacted performance, in cases to a degree that offset the addition of team members. For example, when we studied whether adding a third team member as a supervisor improved team performance, we found that the same performance gains as engendered by the addition of the team member could be observed when the original UGV operator in a UGV-UAS team had very high spatial abilities [12]. In this case, the team member with the high spatial abilities could lead the team to great performance, and the addition of a third team member added little in performance. Similarly, team performance shifted by amounts similar to those related to the addition of a team member when we changed the type of communications and interactions that the operator team could perform. For example, allowing teams of two operators of dissimilar vehicles to share imagery significantly and substantially improved their performance in RSTA tasks over solely being allowed to communicate verbally or via text. Allowing the team members to share not only imagery, but also control over their partner's reconnaissance payload, however, was associated with reductions in performance as the team members got in each other's way and unfamiliarity with the other systems sensors made individual performance poorer [13].

Another variable that interacted with the team members' individual abilities, as well as with the technical and procedural solutions they were given to use, was the LOA and automation of the UVs. In a replication of a previous study, we found that different aspects of the RSTA task benefited from different levels of UV autonomy. For example, with respect to situation and spatial awareness of the operator(s), we found that a middle LOA, similar to supervisory control in automation, was most beneficial. In this case, operators had the highest spatial awareness when the system could traverse point-to-point without needing constant operator intervention (i.e., did not require teleoperation), but required the operator to actively select from system-generated suggestions and options for the next waypoint, rather than executing the movement fully autonomously. Conversely, when the task that was automated required substantial interaction with the system(s), such as inputting commands and waypoints to two vehicles so that they could execute a joint task, a higher LOA that addressed those specific tasks was helpful and associated with better performance.

Finally, the series of studies clearly demonstrated that our conceptualization of the three components of performance for successful human-robot teams was correct. Understanding the RSTA performance of teams of multiple operators interacting with multiple UVs, required an understanding of (a) planning and shared understanding of the team members, (b) spatial and situation awareness, and (c) target recognition and identification. When teams were constrained or deficient in any of the three underlying aspects, their team performance would suffer. Also, attempts to improve the performance of teams would have to target specific variables influencing performance on the underlying topics. For example, spatial abilities of the team members influenced not only spatial and situation awareness but also mission planning and ultimately team performance. Consequently, it is important to assess the impact of a system change or intervention not only on the most closely related underlying aspect, but also the others, when one wants to improve overall team performance.

3.2 Guidelines

1. At this point, operator-to-vehicle ratios of less than 1:1 are unrealistic in RSTA tasks. Although a single operator may be able to control more than one unmanned system in other tasks, such as logistics (“mule-train”) or casualty evacuation, this is not the case for RSTA tasks where creating small teams of operators that interact at 1:1 operator-to-vehicle ratios or higher is beneficial.
2. LOAs should be chosen to match the appropriate task aspect. For spatial awareness, neither manual control (i.e., teleoperation) nor full autonomy are desirable; the former requires too many operator cognitive resources, whereas the latter leaves the operator “out-of-the-loop” and therefore unaware of system status and location.
3. Promising targets for automation are task aspects that require high levels of operator involvement and particularly “busywork.” Path planning, for example, is best supported by automation that simplifies the entry of coordinates (e.g., from keystroke entry to map point-and-click) and combines the different task demands in one, integrated display.
4. The unique and frequently unfamiliar views from robot-mounted displays are difficult to interpret for human operators (e.g., “soda straw effect”). This makes target location and identification very difficult. Automation and structured training are needed if operators are not to spend long hours in inefficient on-the-job training.
5. The impact of technical and procedural interventions/changes to the UV system needs to be considered at multiple levels. An intervention that facilitates path planning but lowers SA may do more harm than good. Conversely, providing the

operators with integrated displays that show obstacles, own forces, likely threats, etc. may improve planning, SA, and target detection simultaneously.

6. Sharing information among team members is typically positive; it allows them to build shared mental models and reduces the need for communications to identify and localize features in the environment. Allowing team members to share control of vehicles, however, is typically not beneficial as it interferes with the team members' individual tasks and can lead to confusion.
7. The spatial ability of individual operators is a powerful predictor of individual and team performance in human-robot teams. In some situations, one team member with high spatial abilities can make up for the addition of a team member that would otherwise coordinate the actions of the team.

4 Field Studies at ARL-Ft. Benning

The purpose of the research program at Ft. Benning was to develop intuitive interfaces for the dismounted infantry to control small UGVs (SUGVs) in a realistic environment. Ft Benning, Georgia was ideal because of the supply of experienced infantry Soldiers who gave the experimenters valuable feedback as well performance data expected from target audience Soldiers. The initial efforts involved investigating scalable displays concepts to find the best solutions for a display in terms of size and performance. Subsequent studies investigated controls, multimodal interfaces, goggle-mounted displays, voice controls, LOA, telepresence and the use of cell phones to control the robot. The purpose of the latter studies was to understand th effects of progressive autonomy on Soldier effectiveness and SA. In general, the procedures were similar, with the Soldiers being in ether a mobile or stationary control condition. The multivariate performance data included deviations from the course, time to complete the course, target detection and latency [14].

4.1 Summary of Results

The initial study determined that the operator could drive TALON SUGVs effectively with 3.5 inch and 6.5 PDA sized displays [15]. Goggle-mounted displays caused perceptual problems during the experiment without compensatory performance gains in three experiments [15][16]. It is important to note that dismounted operations will require map and other information for the Soldiers to navigate effectively beyond line of sight in complex terrain. In the next study [17], Redden and colleagues examined three options to combine map and video control: 1) toggle between two 3.5-inch displays, 2) use a 6.5-inch split screen display, or 3) use a 3.5-inch toggle display with a tactile belt to replace the visual map display. They found that navigation with a 3.5-inch display with tactile augmentation performed as well as the 6.5-inch split screen display indicating that a tactile belt was an efficient means of displaying map information while reducing display real estate. In contrast, when Soldiers were required to toggle back and forth between the two 3.5-inch displays, their performance proved to be less effective [18]. Other studies investigated control options for reducing control surfaces [19][20]. The most efficient way to reduce controller size for discrete functions was a voice activated system permitting hands-free operations, an extremely important advantage for the infantry; however, voice control had limited effectiveness when commands required a continuous motion such as turn left [20]. Reducing control size by miniaturizing controls was preferable to reducing the number of controls by creating more multifunction controls [19]. Recent research investigated more advanced systems developed by ARL, TNO laboratories in the Netherlands, and the SPAWAR Naval Laboratory. The ARL experiment investigated the use of the Android size (about 2.5 inches) cell phone to control semi-autonomous PACBOTs resulting in limited success because of lack of control responsiveness [21]. The TNO telepresence experiments showed that adding auditory 3-D augmentation along with a head-track camera control improved robotic navigation time and target identification for audio-based localization tasks, compared to visual displays [22]. The Navy researchers at SPAWAR used a Multiple Operations Control Unit (MOCU) to monitor autonomous and semi-autonomous SUGVs which were compared to teleoperated systems. Participants performed significantly more rapidly and were less error prone for driving performance using the MOCU interface to monitor the autonomous and semi-autonomous robots compared to teleoperated conditions while their ability to find targets was not affected by LOA [23].

4.2 Guidelines and Lessons Learned

1. Displays as small as 3.5 and 6.5 inches can be used for camera-based teleoperation and local surveillance with small, slow-speed robots and displays as small as 4.3 inches are possible for viewing of video from remote UGVs and UASs.
2. Monocular helmet –mounted displays (HMDs) for robots can be problematic and should be carefully assessed before use to reduce the chance that binocular rivalry, over-stimulation, excessive head borne weight and loss of local SA. They also appear problematic for passive viewing, especially when a controller such as a mouse must be used to complete a task.
3. Speech-based control shows promise for robotic operation when commands are intuitive and based upon the target population word use but not for all functions. They are easy to use and show the potential to increase secondary task performance. They have provided decreased time and effort when performing simultaneous tasks in conjunction with manual controls. Continuous control commands (e.g., pan, drive in a circle) are more efficiently controlled manually.
4. Our results show the efficacy of a tactile belt to augment robotic OCUs to reduce required display size for navigation beyond line of sight.
5. Reducing a controller size by shrinking the size of the individual controls had less adverse impact on performance than reducing its size by reducing the number of controls.
6. As levels of autonomy increase, workload, reconnaissance times, and driving errors decrease and accuracy of mental maps increase.
7. Soldiers prefer the bounding technique (moving and stopping to control the SUGV) to the continuous movement technique when moving with robots. Fewer driving and off course errors were made and more items were detected when the bounding movement was used than when the continuous movement was used. Until robots become more autonomous in their navigation, robotic control during Soldier movement is beyond the multitasking ability of most Soldiers. As reliable

autonomy in robots increases, Soldiers should have more available free time to divert their attention away from the robot and toward their own environment, enabling side-by-side movement.
8. Soldiers with higher video gaming skill appear to be more proficient with the hand and eye component of teleoperations.
9. Visual attendance on robotic displays should not be wider than the 5 degree macular vision (the actual width depends upon the viewing distance) in order to avoid visual scanning which can result in fatigue and missed data.
10. Head-track cameras and telepresence sensors were associated with faster, easier performance, and were preferred by Soldiers over baseline joystick controls with mono vision and audio. However, there is still a challenge with telepresence because of motion sickness experienced by some operators.

5 Conclusions

As Soldiers become more involved in supervising and controlling UVs, they will face multiple challenges related to multi-tasking, poor interface design, manual control, automation reliability, deficient skill sets, and poor SA. Twenty-three specific guidelines to overcome these barriers were derived from three diverse lines of HRI research. The guidelines included criteria for: (a) selection and training of Soldiers that are sensitive to individual differences in attentional control and spatial reasoning, (b) automated decision support, (c) use of intelligent agents, (d) concepts to improve teaming and shared mental models, (e) multimodal and scalable interfaces. Furthermore, the results will help establish a framework for the development of progressive autonomy strategies to reduce operator's workload while improving SA.

References

1. Murphy, R.: Human-Robot Interactions in Rescue Robots. IEEE Trans. Sys., Man & Cybern. 34, 1–15 (2004)
2. Chen, J.Y.C., Durlach, P., Sloan, J., Bowens, L.: Human Robot Interaction in the Context of Simulated Route Reconnaissance Missions. Military Psychology 20, 135–149 (2008)
3. Chen, J.Y.C., Joyner, C.T.: Concurrent Performance of Gunner's and Robotics Operator's Tasks in a Multi-Tasking Environment. Military Psychology 21, 98–113 (2009)
4. Chen, J.Y.C., Terrence, P.I.: Effects of Tactile Cueing on Concurrent Performance of Military and Robotics Tasks in a Simulated Multi-Tasking Environment. Ergonomics 51, 1137–1152 (2008)
5. Chen, J.Y.C., Terrence, P.I.: Effects of Imperfect Automation on Concurrent Performance of Military and Robotics Tasks in a Simulated Multi-Tasking Environment. Ergonomic 52, 907–920 (2009)
6. Chen, J.Y.C., Barnes, M.J.: Supervisory Control of Robots Using RoboLeader. In: Human Factors & Ergo. Soc. 54th Annu. Mtg, pp. 1483–1487. HFES, Santa Monica (2010)
7. Chen, J.Y.C., Barnes, M.J., Kenny, C.: Effects of Unreliable Automation and Individual Differences on Operator's Supervisory Control of Multiple Ground Robots. In: 6th ACM/IEEE International Conference on Human-Robot Interaction. ACM, New York (2011)

8. Chen, J.Y.C., Haas, E.C., Barnes, M.J.: Human Performance Issues and User Interface Design for Teleoperated Robots. IEEE Transactions on Systems, Man, and Cybernetics–Part C: Applications and Reviews 37, 1231–1245 (2007)
9. Chappelle, W.L., McMillan, K.K., Novy, P.L., McDonald, K.: Psychological Profile of USAF Unmanned Aerial Systems Predator & Reaper Pilots. Aviation, Space, and Environmental Medicine 81, 339 (2010)
10. Lee, J., See, K.: Trust in Automation: Designing for Appropriate Reliance. Human Factors 46, 50–80 (2004)
11. Jentsch, F., Evans, A., Ososky, S.: Military HRI Research Conducted Using a Scale MOUT Facility. In: Barnes, M., Jentsch, F. (eds.) Human-Robot Interactions in Future Military Operations, Ashgate, Hampshire, UK, pp. 419–431 (2010)
12. Fincannon, T., Evans, A., Phillips, E., Jentsch, F., Keebler, J.: The Influence of Team Size and Communication Modality on Team Effectiveness with Unmanned Systems. In: Human Factors & Ergo. Soc. 53rd Annu. Mtg, pp. 419–423. HFES, Santa Monica (2009)
13. Fincannon, T., Keebler, J., Jentsch, F., Phillips, E., Evans, A.: Team Size, Team Role, Communication Modality, and Team Coordination in the Distributed Operation of Multiple Heterogeneous Unmanned Vehicles. J. Cog. Eng. & Decision Making (Special Issue on Improving Human-Robot Interaction) (in press)
14. Redden, E.S., Elliott, L.R.: Robotic Control Systems for Dismounted Soldiers. In: Barnes, M., Jentsch, F. (eds.) Human-Robot Interaction in Future Military Operations, Ashgate, Hampshire UK, pp. 335–351 (2010)
15. Redden, E.S., Pettitt, R.A., Carstens, C.B., Elliott, L.R.: Scalability of robotic displays: Display size Investigation. Tech. report (ARL-TR-4456), ARL, APG, MD (2008)
16. Oron-Gilad, T., Redden, E.S., Minkov, Y.: Scalable OCUs for the Dismounted Soldier: Utilizing Unmanned Vehicles – A Field Study. J. Cog. Eng. & Decision Making (Special Issue on Improving Human-Robot Interaction) (in press)
17. 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. Tech. report (ARL-TR-4708), ARL, APG, MD (2009)
18. Redden, E.S., Elliott, L.R., Pettitt, R.A., Carstens, C.B.: A Tactile Option to Reduce Robot Controller Size. J. Multimodal User Interfaces 2, 205–216 (2009)
19. Pettitt, R.A., Redden, E.S., Carstens, C.B.: Scalability of Robotic Controllers: An Evaluation of Controller Options. Tech. report (ARL-TR-4457), ARL, APG, MD (2008)
20. Pettitt, R., Redden, E., Carstens, C., Elliott, L.: Scalability of Robotic Controllers: Speech-based Robotic Controller Evaluation. Tech. report (ARL-TR-4858), ARL, APG, MD (2009)
21. Pettitt, R., Redden, E.S., Fung, T.: Scalbility of Robotic Controllers: An Evaluation of Controller Options, Experiment II. Tech. report. ARL, APG, MD (in review)
22. Elliott, L.R., Jansen, C., Redden, E.S., Pettitt, R.: Robotic Telepresence: Perception. Performance, and User Experience (in review)
23. Pettitt, R., Redden, E.S., Pacis, E., Carstens, C.B.: Scalability of Robotic Controllers: Effects of Progressive Levels of Autonomy on Robotic Reconnaissance Tasks. Tech. report (ARL-TR- 5258), ARL, APG, MD (2010)

Optimizing Performance Variables for Small Unmanned Aerial Vehicle Co-axial Rotor Systems

Jonathon Bell, Mantas Brazinskas, and Stephen Prior

Middlesex University, School of Engineering and Information Sciences, Trent Park Campus,
Bramley Road, London N14 4YZ, United Kingdom
Jonathon.bell@hotmail.com

Abstract. The aim of this project was to design and build a test-rig that is capable of analyzing small unmanned aerial vehicles (SUAV) co-axial rotor systems. The intention of the test-rig development was to highlight important aeromechanical components and variables that dictate the co-axial units flight performance, with the intention of optimizing the propulsion systems for use on HALO® a co-axial SUAV designed by the Autonomous Systems Lab at Middlesex University. The major contributions of this paper are: an optimum COTS co-axial configuration with regards to motor and propeller variations, a thorough review and validation of co-axial rotor systems inter-rotor spacing which in turn identified an optimum H/D ratio region of between (0.41–0.65).

Keywords: Co-axial Rotor, SUAV, Aerodynamics, H/D ratio.

1 Introduction

This paper details the background, concept and investigations into co-axial rotor systems used on full-scale helicopters through to Micro Air Vehicles, with the intent to highlight the key aerodynamic and aeromechanical components which contribute to the systems performance in the flight condition of hover.

The contra-rotating co-axial rotor design offers many advantageous attributes over single rotor systems, with the most often cited advantages being the reduction of the overall rotor diameter of the co-axial rotor system, and lack of need for a traditional tail rotor (which has been estimated to consume 5-20% of the total power produced). These areas are accentuated and highlighted when the design and optimisation of co-axial rotor system at the scale of small UAVs, which also requires a greater understanding of the performance variables that affect the co-axial propulsion system at low Reynolds Number (Re) operation, are investigated.

Recent co-axial rotor research relies heavily upon outdated co-axial rotor system studies, theoretical modelling, and computational fluid dynamics. There is very little empirical data and evidence outside the report of Coleman [1], together with research commenced by a select few research and development departments at universities across the world that identify the optimum conditions of co-axial rotor systems, especially at the SUAV scale. Even with the current research and data available it is difficult to predict the performance and optimize a co-axial rotor system for a specific scale due to conflicting reports.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 494–503, 2011.

Much of the funding, currently worth an estimated production value of US$ 2.05 billion (2010-19), for the research of SUAVs (which incorporates co-axial rotor systems) is predominantly fuelled by the international military, where the SUAV rotary winged systems are pitched to play increasingly more vital roles in ISTAR (Intelligence, Surveillance, Target Acquisition & Reconnaissance) operations. The project and study of these exotic systems has been closely aligned with the co-axial tri-rotor small UAV, HALO™ which is in development within the Autonomous Systems Laboratory at Middlesex University.

2 Co-axial Rotor System Aerodynamics

As aerodynamics and aeromechanics have the greatest influence on SUAVs in-flight performance, this section is a summation of the core components that influence the co-axial rotor system in the flight condition of hover, and in turn have influenced the testing variables used during the analysis phase. Although the evaluation of forward flight is of interest, it was deemed too complex with respect to fabricating a controlled environment such a wind tunnel to be able to simulate these conditions and was considered unfeasible within the constrictions of the project time limit.

The Figure of Merit (FM) when applied to a co-axial rotor system is a non-dimensional efficiency metric that provides a basis to conduct a relative comparison of rotor performance. The FM uses the "ideal" power required to hover (calculated using the moment theory) that is in turn equated against the "actual" power required to hover. An equation for the Figure of Merit by Leishman [2] is given as follows:

$$FM = \frac{1.2657 \frac{C_{T_l}^{\;3/2}}{\sqrt{2}} \left[\left(\frac{C_{T_u}}{C_{T_l}} \right)^{3/2} + 1 \right]}{K_{\text{int}} K \frac{C_{T_l}^{3/2}}{\sqrt{2}} \left[\left(\frac{C_{T_u}}{C_{T_l}} \right)^{3/2} + 1 \right] + \frac{\sigma C_{d_o}}{4}} \tag{1}$$

In terms of the measured co-axial systems power, the definition for FM is:

$$FM = \frac{1.2657 \frac{C_{T_l}^{3/2}}{\sqrt{2}} \left[\left(\frac{C_{T_U}}{C_{T_l}} \right)^{3/2} + 1 \right]}{C_{P\,meas}} \tag{2}$$

Where:

Ctu + Ctl	=	Rotor Thrust coefficient (Upper, Lower)
Cpmeas	=	Rotor Power coefficient measured
σ	=	Rotor solidity
Cdo	=	Minimum or zero-lift drag coefficient

Rotor flow fields discussed by Leishman and Ananthan [3] are referred to as the *vena contracta* of the upper and lower rotors; it is also referred to as the slipstream of the co-axial rotors. To minimize the interference-induced power factor using the momentum theory the co-axial rotor system is theoretically set in a condition of “the rotors operating at balanced torque, with the lower rotor operating within the *vena contracta* of the upper rotor”[4] as discussed below. Leishman goes on to discuss the ideal flow considerations noting that “one-half of the disk area of the lower rotor must operate in the slipstream velocity induced by the upper rotor” [3]. The flow model of a co-axial rotor system and the vena contracta are detailed in Figure 1.

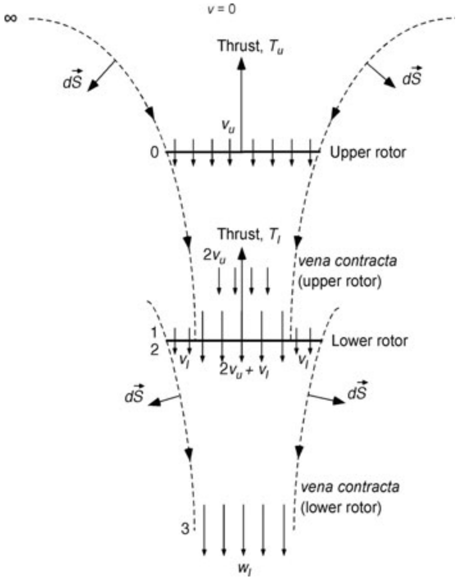


Fig. 1. Flow Model of a Co-axial Rotor System [4]

The separation distance could therefore have an effect upon the severity of the interference-induced power loses, which would in turn possibly increase the efficiency rating (FM) of the co-axial rotor system.

2.1 Testing Variables

The investigation of the co-axial rotor system primarily revolved around four testing variables, with the aim of this paper focusing on the results on co-axial inter-rotor spacing & system configuration:

- **Inter-rotor spacing** – The separation distance (H) between the co-axial rotor system discs. Inter rotor spacing is one of the fundamental components of the SUAV co-axial system which has been tested due to the associated aerodynamic effects; interference-induced power losses, wake contractions, and rotors *vena contracta*. The H/D ratio is used as a non-dimensional figure to enable comparison of multiple systems across a range of scales.

Figure 2 compares H/D ratios, incorporating full-scale co-axial helicopters to MAVs. The table demonstrates that the SUAV example systems have a significantly higher H/D ratio (average H/D = 0.315), when compared with the average for full-scale systems having an H/D = 0.09.

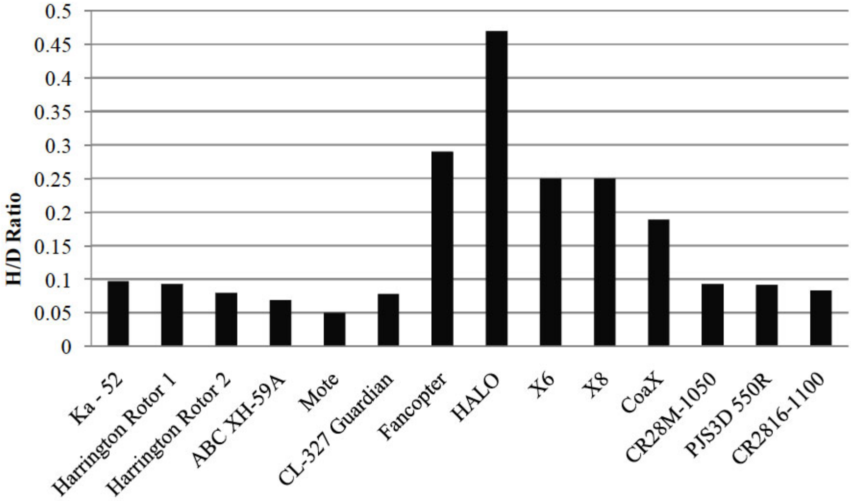


Fig. 2. Inter-rotor Spacing Comparison Chart [5]

- **Propeller Pitch** - The propellers used in the co-axial tests are fixed pitch, but unlike full-scale rotor blades that have an almost uniform pitch throughout the diameter due the design preference of a symmetrical blade section [6]; the test propellers have a varying pitch.

- **Propeller Diameter** (upper and lower) - The diameter of a propeller is one of the most important characteristics in determining the induced power of a rotor system:

$$Pi = \frac{T^{\frac{2}{3}}}{\sqrt{2\rho A}} \tag{3}$$

It has been shown in studies by Leishman that the larger the rotor diameter the lower the disc loading, induced velocities, and a decrease in induced power requirements [2]. Andrews [8] notes that an 8% reduction in upper rotor radius enhances the performance of the lower rotor due to an increase of exposed clean air. This variable will be controlled only using a select ‘family’ of propellers (rotors) to determine the performance attributes related to the decrease of the upper and lower rotors.

- **Co-axial Rotor Configuration** – The co-axial propulsion unit tested has individual motor units powering the upper and lower rotors. This allows for multiple variations and configurations of the orientation of the motors and propellers to be analysed and the results recorded respectively.

3 Test-Rig Development

Recent developments in the co-axial rotor system for the small-scale UAV sector have resulted from the technological advances in RC propulsion units [7]. One of the earliest recorded co-axial UAV studies was work commenced by Andrews [8] on a Westland Helicopter Ltd developed system called Mote, the system's handling and control qualities are discussed in detail by Faulkner and Simons [9]. It was these studies by Andrews [10] that demonstrated a decrease of 8% to the upper rotor radius enables "the enhanced performance of the lower rotor as proportionately more disc is exposed to clean air". Andrews also discussed the inter-rotor spacing stating that there are no "practical" gains after H/D = 0.05.

More recent test-rigs and co-axial rotor system investigations include the work of the Autonomous Systems Lab (ASL), ETH at Zurich. Bouabdallah has spearheaded the extensive work produced by this team [11]. The significant research systems/platforms developed by the ASL at ETH are CoaX and CoaX 2, both co-axial MAV's. Unlike many co-axial rotor studies the muFly team has designed and built their own co-axial rotor test bed, and recorded the study in detail. A similar system that enables the investigation of MAV co-axial rotor systems is the rig developed by the University of Maryland for the MICOR MAV. Both systems are designed for variable pitch rotor heads.

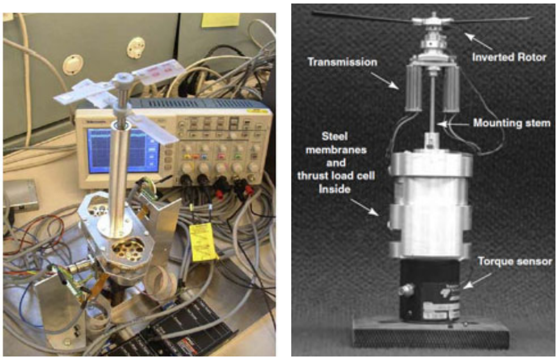


Fig. 3. muFLY [11] & UMD MICOR co-axial rotor system test-rigs [12]

The test-rig's priority was to be able to test and measure various co-axial fixed-pitch rotor system configuration variables. The components used in the setup for a co-axial rotor system (using HALO's components as a datum) have dictated the majority of the test-rigs overall design. The motors used for the co-axial rotor system are the AXI 2217/20 electric Outrunner DC motor, which are inherently stable and give good efficiency ratings of approximately 82%. The propellers used range from dual-bladed, low pitch and slow fly APC 10 inch propellers up to 12 inch Master Airscrew tri-bladed propellers. The range tested encompasses five ‘families’ of propellers, each with their own performance benefits.

Taking into account the co-axial rotor systems testing variables, and the known datum components set by the HALO configuration, mechanical solutions were developed. Linear motion technology in the form of a motor driven lead-screw was chosen for the inter-rotor spacing control of the co-axial rotor configurations. The desired range of inter-rotor spacing stemmed from using the GWS 1060X3 propeller as a datum measure (10 inch or 254 mm). This permitted the H/D range to be varied within the range (0.08–1.0).

The optimization process of the co-axial rotor system was continually taken place as the testing commenced. To develop a portfolio of test data from the testing components, analyze the efficiency of particular component configurations and testing conditions a data logging and live monitoring tool has been employed. The Hyperion Emeter II is a high performance measurement tool that is able to measure, analyze, and log key performance factors used in electric systems and RC models. The Emeter is supplied with a remote data unit (RDU) which houses a high precision shunt that is capable of accurately handling high currents and voltages, and is able to feed this data back to the Emeter for evaluation purposes.

4 Analysis and Results

To be able to test the co-axial configuration in the optimal motor and propeller arrangement a series of tests containing various co-axial configurations were analyzed. Eight configurations were used for the optimal motor and propeller configuration for a co-axial propeller system, with only four having contra-rotating rotors. A comparison data set consisting of individual rotors at multiple orientations used in the co-axial configurations gave a datum result for each singular rotor’s performance.

The highest performing co-axial configuration, when plotting the measured system Thrust (g) Vs Speed (RPM x 1000), was when the motors are placed on the outside of each mounting arm on the test-rig using an upper – Pusher propeller, and lower – Tractor propeller setup. A similar overall performance measurement was seen when plotting Output Power (W) Vs Speed (RPM x 1000). This data coincides with the finding of Shkarayev [13], where the rotor configuration used on the SUPAERO MAV showed a 20–23% thrust increase when using a pusher configuration when compared to a tractor configuration.

As co-axial rotor systems are compared to their singular counterparts in numerous studies, a study of the individual rotor and motor configurations used in the co-axial testing has also been undertaken. The points of interest and observations are detailed below:

- When comparing the co-axial rotor configurations measured Thrust against the combined two singular rotor systems measured Thrust, the average Thrust output is 23.15% lower.

- The Thrust/Current Ratio of the co-axial rotor system averages a 2.22% decrease per Ampere when compared to the combined singular Rotors.
- Independently the individual tests of each singular rotor comparison gave unexpected and interesting results. Prior to the experimentation phase it was thought that a tractor and pusher propeller operate in an identical fashion, i.e. producing similar Thrust, and Output Power performances (allowing for the inaccuracies of the test-rig, and data logging). Figure 4 depicts the performance variation of the Tractor and Pusher GWS 1060X3 HD propeller in two configurations for each type of propeller. The pusher propeller placed on the upper arm had a thrust increase (at 7,000 RPM) of 7.11% compared to the Tractor Propeller; this trend was also observed on the lower rotor comparison, with the Pusher variant producing 8.29% (at 7,000 RPM) more thrust than its tractor counterpart.

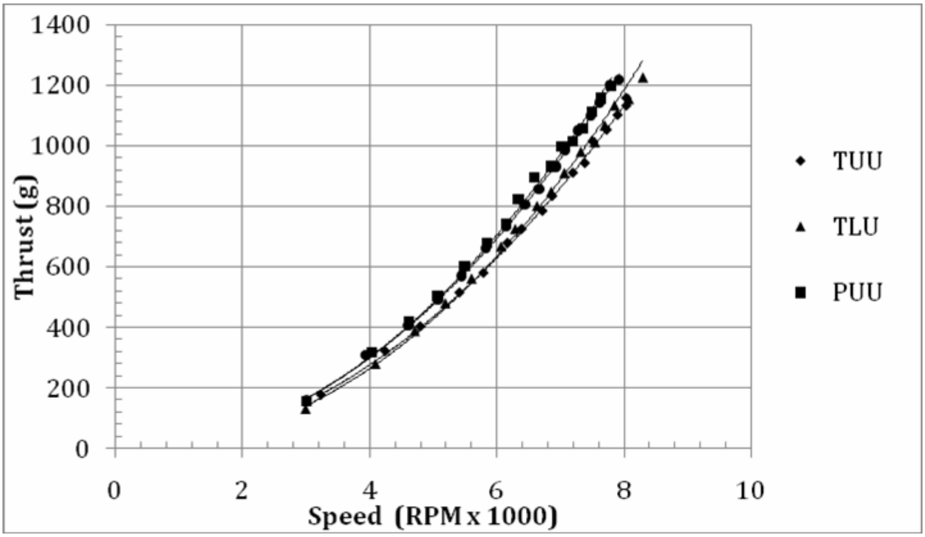


Fig. 4. Individual Motor Configurations - Comparison of Tractor and Pusher Rotors

Using the optimally determined configuration for the co-axial rotor system, inter rotor spacing tests were commenced with a range from 20 mm to 250 mm (0.08 < H/D < 1.0) at 10 mm increments. The system was operated at an unequal torque and thrust balance, with the objective of the testing to establish a co-axial rotor systems static thrust capabilities at a given H/D ratio. As the research is to coincide with the development of the ASLs' HALO™ SUAV the propeller and motor combination of primary focus was the GWS 1060X3 HD and the AXI 2217/20.

Figure 5 is a select region of H/D ratios that provided a measurable increase in Thrust at a given Current (A). A range of 12–14 A was used to plot the variation in Thrust Vs H/D ratio, with an H/D ratio of approx. 0.5 showing the highest thrust.

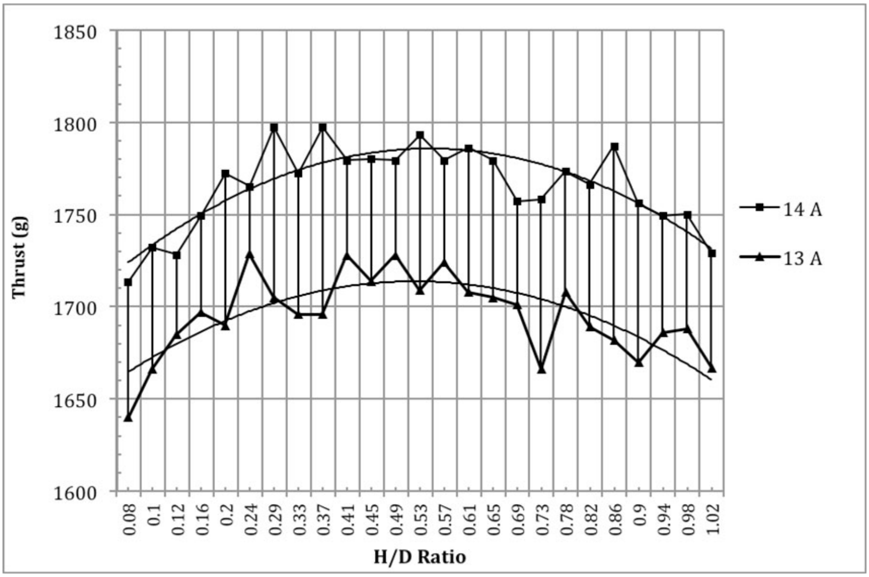


Fig. 5. Variation of Co-Axial Thrust with H/D Ratio

5 Conclusions and Future Work

There have been multiple areas explored in the process of optimizing a SUAV co-axial rotor system, some of which have had limited research exposure and others which have been detailed thoroughly.

One of the main areas of interest and which has had the greatest influence on the co-axial tests-rig design was the inter-rotor spacing attribute of the co-axial rotor system. The H/D ratio has been prominent in many significant papers, but lacking an empirical value or an optimal dimensionless condition. In this paper the H/D ratio of a SUAV has been explored thoroughly, reviewing the systems performance at incremental stages, the findings from this study have shown that a range of H/D ratio of between (0.41–0.65) is advantageous in the performance of SUAV systems. This finding lends itself to the theory of inter-rotor spacing is a non-dimensionally similar figure, which cannot be applied across a spectrum of systems; this could be attributed to the viscous losses of flight at low Reynolds Numbers (< 50,000).

5.1 Test-Rig Review

The foundation of the optimization process for the co-axial rotor system was the design and development work of the co-axial test-rig. The system was designed to cater for the requirements and variables that were initially deemed to cover all the testing attributes of a Small Unmanned Aerial Vehicle co-axial rotor system. Although the test-rig was able to cater for the fundamental components of the testing process it did however lack mechanisms and testing apparatus that would have in hindsight allowed for greater and more in-depth analysis of the co-axial rotor system, especially highlighting the individual motor performance within the co-axial unit.

Current research within the Autonomous Systems Laboratory at Middlesex University involves the design and development of the Mark II co-axial test-rig. As briefly mentioned previously the test-rig is being designed to analyse some of the key attributes of the co-axial system that had been overlooked in the original test-rig.

A critical appraisal of the original test-rig and an indication of future improvements are stated below:

- One of the failings of the original test-rig was the lack of a real-time reaction torque sensor. Due to this lack of component it was difficult to measure and interpret the co-axial rotor systems yaw torque balance. As the testing process developed the need for the inclusion of this sensor became apparent.
- Individual rotor thrust is calculated using the thrust constants and factors from the individual rotors static performance graph. For future work and developments to the test-rig, the design should incorporate individual load cells. This key attribute would enable a complete assessment of the operating conditions of the upper and lower rotors independently, and thus provide insight into the induced loses of the co-axial rotor system.
- The future test-rig may incorporate an automated control and recording system such as NI LabView (for data acquisition and test bench control). When employed for further testing this would provide greater accuracy, data analysis and simulation possibilities.

References

1. Coleman, C.P.: NASA Technical Paper 3675, 32 (March 1997)
2. Leishman, J.G.: Principles of Helicopter Aerodynamics. Cambridge University Press, USA (2002)
3. Leishman, J.G., Ananthan, S.: Aerodynamic Optimization of a Coaxial Proprotor. In: Annual Forum Proceedings - American Helicopter Society, vol. 62(1), pp. 64–86 (2006)
4. Leishman, J.G., Syal, M.: Figure of Merit Definition for Coaxial Rotors. Journal of the American Helicopter Society 53(3), 290 (2008)
5. Bell, J.: Investigations into Optimal Co-Axial Rotor System Configurations for Small UAVs (Published Masters Thesis). Middlesex University, UK (2010)
6. Lakshminarayan, V.K.: DRUM: Computational Investigation of Micro-Scale Coaxial Rotor Aerodynamics in Hover (Published PhD Thesis). University of Maryland, USA (2009)
7. Prior, S.D.: Reviewing and Investigating the Use of Co-axial Rotor Systems in Small UAVs. International Journal of Micro Air Vehicles 2(1), 1–16 (2010), doi:10.1260/1756-8293.2.1.1
8. Andrews, J.M.: Coaxial Rotor Aerodynamics (Published PhD Thesis) Southampton University, UK (1981a)
9. Faulkner, A.J., Simons, I.A.: The Remotely Piloted Helicopter. Vertica 1(3), 231–238 (1977)

10. Andrews, M.J.: Coaxial Rotor Aerodynamics in Hover. Vertica 5, 163–172 (1981b)
11. Schafroth, D., Bouabdallah, S., Bermes, C., Siegwart, R.: From the Test Benches to the First Prototype of the muFly Micro Helicopter. Journal of Intelligent and Robotic Systems 54(1-3), 245–260 (2008)
12. Bohorquez, F.: DRUM: Rotor Hover Performance and System Design of an Efficient Coaxial Rotary Wing Micro Air Vehicle (Published PhD Thesis). University of Maryland, USA (2007)
13. Shkarayev, S., Moschetta, J., Bataille, B.: Aerodynamic Design of VTOL Micro Air Vehicles. In: MAV 2007 Proceedings, France (2007)

Trust Evaluation through Human-Machine Dialogue Modelling

Cyril Crocquesel[1,2], François Legras[3], and Gilles Coppin[1,2]

[1] Institut Télécom; Télécom-Bretagne; UMR CNRS 3192 Lab-STICC, France
[2] Université Europénne de Bretagne, France
[3] Deev Interaction, France

Abstract. Trust in automation, and particularly maintaining an adequate level of trust in automation is now recognized as a major performance factor in supervisory control. Leveraging man-machine interaction is seen as a promising approach to influence the level of trust of an operator. Two problems need to be addressed in order to reach this goal: first measuring the level of trust; second acting on the level of trust to reach a more appropriate level. In this paper, we tackle the first problem, and propose to use a computational dialogue modelling approach to evaluate trust dynamically. We describe our model on two examples and give some perspectives.

1 Trust in Automation

Supervisory control [15] raises the question of interfacing human operators and automata, which have become more and more autonomous and complex along time. This increase in complexity highlights the necessity of establishing an adequate relationship between the operator and the system, a relationship that goes beyond the pure management of performance and addresses the question of trust as well. The degree of trust that a human operator gives to automation is now recognized as a major factor in the success (or failure) of such systems [13,7].

Two classes of trust-related problems are identified in the literature [13]:

- misuse stems from over- or under-reliance on automaton. The capabilities of an automated system can be badly appreciated by an operator. Thus, an under-estimation of a system's capabilities or a non-adapted operator trust (too much or not enough) can lead to erroneous uses and negative consequences, e.g. errors, high workload.
- disuse occurs when the operator is so mistrusting of the system that he/she makes no use of a potentially useful automation.

Obviously, trust and mistrust can lead (through disuse and misuse) to severe drawbacks in performance, which human operator and automata separately would not cause, but which interfere mutually and spoil the global functioning of the system. Therefore trust needs to be adequate, in relation to the automata's capabilities.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 504–513, 2011.

In order to reach this level, in our opinion, a dynamic adaptation of the interaction is required. Indeed, we assume that this adjustment can change the status of operator's trust. For this, we must solve two problems. The first involves knowing how to act to influence the degree of trust of the operator. In others words, what are the interaction strategies which impact operator's trust. The second problem is to determine the current state of trust in order to enhance operator trust. So it is necessary to be able to measure the degree of operator's trust. It is this point that we deal with in this paper.

2 Trust Measurement

2.1 Parametric Estimation of Trust

Bhattacharya identifies a certain number of characteristics to define trust [3], which are summed up in a definition by Lee and See [9]: *"the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability"*, the agent being a human or artificial agent. In view of this definition, trust depends of many parameters:

Environment and context: the complexity and uncertainty of an environment will determine the perceived risk and the need of trust [10,3]. So the tendency of a person to trust will depend on his/her individual, cultural and organisational context [9]. The individual context refers to the history of interaction between the trusted and the trustor;

Decision making: trust-based decisions are made when future event are too complex to be predicted. In this case, trust is equivalent to a perceived risk, and the associated subjective probability is the trust degree in the predictibility of the trusted [3];

Outcomes and consequences: individuals not only give probabilities to future actions, but also assign a probability to the desired outcome of the action. This prediction represents the importance attached by the individual to the achievement of a specific outcome and of its consequences (success or failure);

Information about the trusted: Bhattacharya describes trust as multidimensional. These dimensions include in particular the ones defined by Barber [2]: technical competence (skill), persistence (integrity) and fiduciary responsibility (benevolence).

Although these parameters are sufficient to build a static model of trust decision (e.g. see the work of Sutcliffe [16]), Muir [11] leveraged three parameters (issued from Rempel's work [14]) in order to describe the dynamics of trust:

Predictibility plays an important role at the beginning of a trust relationship. The evolution of trust is influenced by the complexity of the automation and the stability of the environment. So, operators must rely on their training and prior experience;

Dependability intervenes during a mature relationship. Trust is more an evaluation of the overall experience than an evaluation of a specific behaviour.

The operator's assessment of the automaton is enhanced when the operator pushes it beyond its design limits;

Faith represents a fully mature trust relationship. The appropriate perception and flexibility of the automaton by the operator allows him/her to control the system effectively without a complete understanding of the automaton's behaviour.

This approach to trust, based on interpersonal trust relationships, enables defining a set of guidelines for the conception of systems which promote an adequate trust degree [12], to guide the training of future operators of given systems [12], or to elaborate a posteriori evaluation questionnaires [8]. So it would not be possible to build an adaptive interaction strategy based on this one.

2.2 Dynamic Evaluation

In 2004, Lee and See conducted a review of trust in automation [9] that led to a descriptive model of trust. Their model shows the influence of individual and environmental contexts on the evolution of trust and indirectly on the operator's intentions. Models of trust are based on an analysis of all input parameters (context). Additionally, Lee describes trust as a closed loop system incorporating the interaction between the operator and the automaton.

Indeed, according to Ajzen and Fishbein [1] and as represented by Lee, trust is an attitude that is one of the precursors of an operator's intentions. These intentions either lead to reliance action or they do not. Therefore, we assume that it is possible to evaluate the trust level of an operator from his/her interaction with the automaton. Current models of trust are based on a posteriori interviews because of the difficulty of dynamically evaluating the operator's perception of risk, or the predictibility he/she attributes to the automaton. Our approach will allow a dynamic assessment of trust. This leads to the possibility of adapting the interaction between operator and automaton, allowing the later to dynamically tune its strategies to enhance the trust level of the operator.

3 Dialogue for Supervisory Control

Grounding theory describes the mechanisms used during the dialogue to construct and maintain the *common ground* i.e. the set of knowledge common to interlocutors [5]. In order to modelize the dialogue in the context of supervisory control — with a focus on monitoring (function of supervisory control [15]) — we adapt Traum's computational model of grounding theory [17]. We present our computational model on two examples.

3.1 Unmanned Air Vehicle Supervision

An operator in charge of an unmanned air vehicle (UAV) has to patrol around a given area. For this, he interacts with the UAV management system through its user interface.

The monitoring function of supervisory control involves the observation of the performance of goals to ensure that they are carried out properly [15]. For this, the system, the initiator 'I' of the dialogue, communicates a piece of information to the operator, the responder 'R'. Each piece of information is attached to a discourse unit (DU), which we represent using a transition network (TN), as illustrated in fig.1.

The transitions describe dialogue acts as:

- **initiate:** introduces the initial information;
- **ack:** acknowledges;
- **repair:** updates the information;
- **cancel:** abandons the initial information sharing;
- **reqext:** asks more informations concerning the initial information;
- **reqexpl:** asks an explanation about the initial information;
- **ext:** answers to *reqext*;
- **expl:** answers to *reqexpl*;

The states of the TN describe the dialogue state:

- **S:** DU hasn't been initiated yet;
- **1:** initiative with I and no discourse obligations (beyond those for R to acknowledge);
- **2:** initiative with I and a discourse obligation for I to repair;
- **3:** initiative with R and no discourse obligations (beyond those for I to acknowledge);
- **F:** DU has been grounded, the attached information is shared in the common ground;
- **D:** DU has been abandoned.

Let us look at an example of a TN routing.

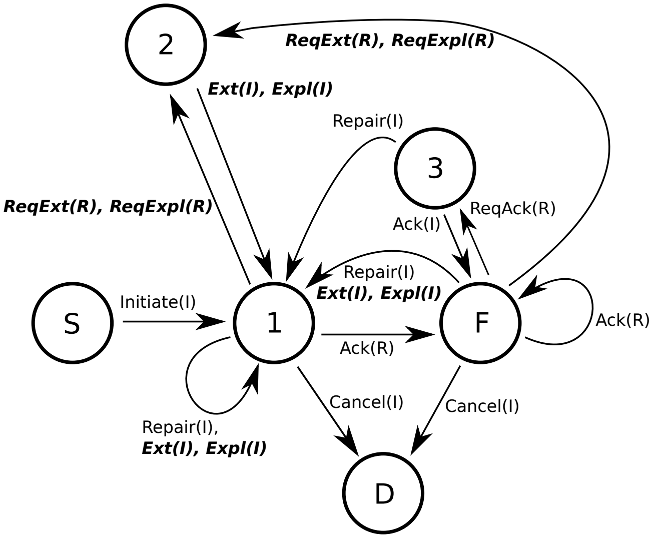


Fig. 1. Transition network adapted to monitoring. In boldface, transitions explaining control mechanisms (explain, extended information).

Fuel level alert. An UAV is currently patrolling to secure an area. Its fuel has reached a low level. The system informs the operator of the low fuel level, so the attached TN switches from state *S* to *1*. This change of state is carried out by the *initiate(I)* transition that corresponds to the presented information. When the operator acknowledges ("OK button) the information, the transition *ack(R)* switches the TN to state *F*. A "check" function attached to the information is available to the operator. With this, he/she can request the information confirmation from the system. The network, following the transition *reqack(R)*, switches to state *3*. The fact that the system refreshes the information causes the TN to switch to state *1* with *repair(I)*, a message validating the information (a pop-up "information checked") makes the network switch to state *F* with ack(I).

Speed of UAV. Some kind of information such as speed is continually displayed and refreshed, in such cases the TN loops on state *1*. An acknowledgement of the information to validate the *ack(R)* transition and to make the network switch to state *F* could be obtained by visual checking with an eye-tracker. The refreshing of the speed by the system reopens the DU and switches the network into *1* state.

3.2 Control within Dialogue

Supervisory control is a control of information issued from the system by the operator. It is important to make the mechanisms for explanation and extended information requests appear explicitly in our dialogue model (fig.1).

Let us take the fuel level alarm example. The current state is *1* and the operator becomes aware of the information. From the point of view of interaction, two functions are attached to the information:

- Explain function: the operator requests that the system explain the alarm. The TN switches to state *2* with *reqexpl(R)* transition. The system answers by a pop-up "fuel level < 5%" that switches the network to state *1* with *expl(I)* transition.
- Extend function: the operator requests that the system display extended information. The TN switches to state *2* with *reqext(R)* transition. The system answers by a pop-up "fuel level 2%" that switches the network to state *1* with *ext(I)* transition.

After the operator has acknowledged the initial information, he could still request explanations and/or extended information. So these mechanisms can be activated from state *F*.

3.3 Information Representation

Requests for explanation and extended information need an information representation which make it possible to build answers for them. For this we decided to establish two graphs: one to represent the explanation links between information, and the second to represent the extended information links.

Explanation graph. Take for example the course information of an UAV which has to reach a waypoint and while having a low fuel level. Each node of the graph is a piece of information. The oriented edges mean "explain" (fig.2). Thus one can read: "command go to" "explain" "course".

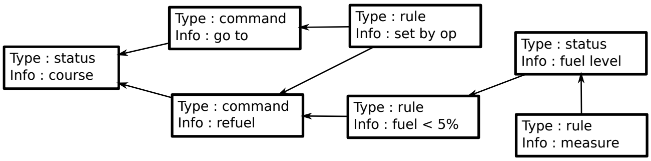


Fig. 2. Example of explanation graph

If we return to our dialogue model, the "course" information is displayed to the operator (fig.4). An "explain" button is attached to it and allows the operator to activate *reqexpl(R)* transition and to switch the network to state *2* (fig.5). The system searches in the explanation graph and selects the right answer: "command refuel" instead of "command go to". The TN switches to state *1* with *expl(I)* transition (fig.6). This answer is due to the fact that "command refuel" has an higher priority than "command go to" for the UAV safety.

A new explanation within the dialogue unit deals with "command refuel" information and not "course status".

Extended information graph. With the same example, let us look at the extended information. Each node corresponds to a piece of information. The oriented edges mean "extend" (fig.3). One can read: "wind speed status" "extend" "course status".

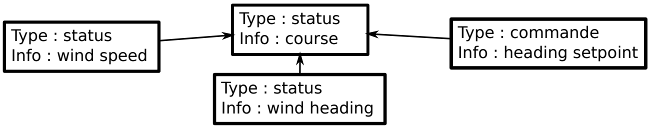


Fig. 3. Example of extended information graph

If we return to our dialogue model, the "course" information is displayed to the operator (fig.4). An "extend" button is attached to it and allows the operator to activate *reqext(R)* transition and to switch the network to state *2* (fig.5). This button is coupled with a list of choices (one, average, complete) which determines the number of pieces of information desired by the operator. The system searches in the extended information graph and selects information depending on the context. Thus, for an average extension, the system answers: "wind speed status" and "wind direction status". The TN switches to state *1* with *ext(I)* transition (fig.7). For a complete extension, the system adds "course setpoint" to the answer.

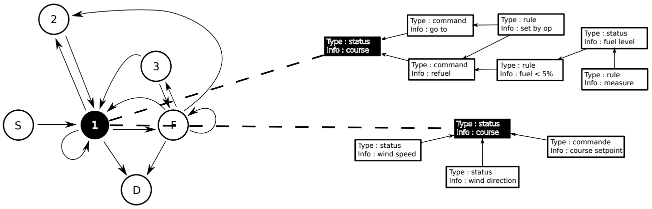


Fig. 4. Interconnection between TN, explanation graph and extended information graph. In state *1*, the TN is linked to introduced information.

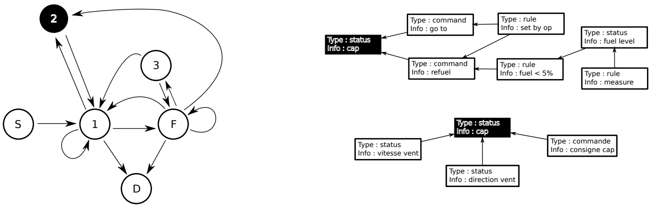


Fig. 5. Interconnection between TN, explanation graph and extended information graph. In state *2*, the introduced information is always activated but not linked to *2* state.

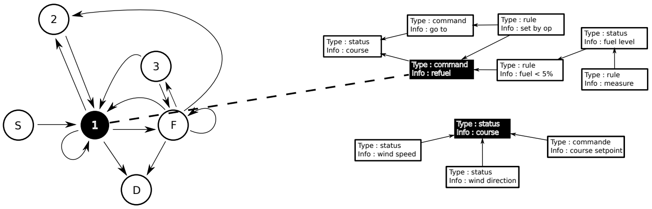


Fig. 6. Interconnection between TN, explanation graph and extended information graph. In state *1* after an explanation request, the TN is linked to explanation information.

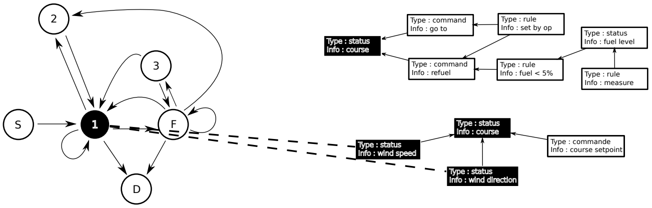


Fig. 7. Interconnection between TN, explanation graph and extended information graph. In state *1* after an extended information request, the TN is linked to extended information.

4 Control and Trust Evaluation

The evaluation of operator's trust from an interaction analysis perspective requires the establishment of a link between dialogue and trust, or more specifically between control and trust. Control is defined by Castelfranchi [4] as an action:

- aimed at ascertaining whether another action has been successfully executed or if a given state of the world has been realized or maintained (feedback or checking);
- aimed at dealing with possible deviations and unforeseen events in order to positively cope with them (intervention).

4.1 Trust/Control Link

Relying on Castelfranchi's definition of broad notion of trust (that qualifies the trust in an agent for a given task in a given environment) [4], we assume that the form of the dialogue between an operator and a system represents, at least indirectly, the level of control he/she wants to apply and consequently the level of trust/mistrust in the system's capabilities.

4.2 Behaviour Hypotheses

We think that our dialogue model is a stochastic process for a given degree of trust. We suppose it is a Markov chain in which the probabilities attached to transitions are specific to the degree of operator's trust. From this, we formulate some hypotheses about the operator's behaviour and its effect on transition probabilities of the Markov chain.

Trusting behaviour. *A trusting operator minimizes the number and frequency of his controls.*

Firstly, if we consider an UAV operator and his/her behaviour as trusting, the transitions *reqexpl(R)* and *reqext(R)* will be rare, their attached probabilities tend to 0.

Secondly, for informations whose content is regularly refreshed (UAV speed status), if the operator is trusting, he/she will seldom check the information that implies a low occurency of the state F in the dialogue sequence. This means the probability attached to *ack(R)* may be correlated to the degree of trust: the lower it is, the higher the degree of trust is.

Thirdly, the dialogue *reqext(R)* acts indicate a need of the operator to build an idea of the current situation. But for this, it is not necessary to have a complete view if the system is trusted. However, a too limited view means the operator overtrusts the system.

Mistrusting behaviour. *A mistrusting operator intensifies his/her controls on the system.*

Firstly, a call to *reqack(R)* is the first mark of mistrust. It results into a switch to state *3* of the TN. Coming back to our example of the UAV operator, when an intrusion alarm occurs, the use of the "check" button attached to the alarm information will indicate mistrust by the operator. Successive use of this button means a complete lack of trust in the system by the operator. So the degree of mistrust may be correlated to the probability attached to *reqack(R)*.

Secondly, if the operator is trained, the explanation request means the operator calls system decisions into doubt. Thus a low probability attached to *reqexpl(R)* transition means a beginning of mistrusting operator behaviour. The higher this probability will be, the more mistrusting the operator will be.

Thirdly, the dialogue acts *reqext(R)* indicate the need of the operator to build an idea of the current situation. However, if he/she requests a maximum amount of information, presumably he/she doesn't trust the system. Of course, it is relative to the context. Indeed, this interpretation is right only if the context has not undergone any change.

5 Conclusion and Future Works

We proposed, in this paper, a theoretical approach to trust evaluation based on observation and analysis of man-machine interaction. The key points of this model are:

– Trust evaluation can be done with the link between trust and control. This implies an analysis depending on the context;
– Man-machine dialogue is representative of the control made by the operator in supervisory control. First results obtained during a preliminary experimentation establish a correlation between a posteriori trust evaluation and corrective user command [6]. So by dialogue pattern analysis, trust could be estimated from the interaction.
– Identification of pattern behaviour gives us the observables for trust evaluation.

Future works have to identify observables of trust estimation and to validate them. For this, we intend to conduct experimentations enabling us to compare our approach with older ones like questionnaires.

Acknowledgements

We thank Mrs Aimée Johansen who helped us to write in English. This Research Work was supported by PhD grant funded by Brest Métropole Océane (Research contract PAGICEA 3CCPT813 : 2008-2011).

References

1. Ajzen, I., Fishbein, M.: Understanding Attitudes and Predicting Social Behavior. Prentice-Hall, Englewood Cliffs (1980)
2. Barber, B.: The Logic and Limits of Trust. Rutgers University Press (1983)

3. Bhattacharya, R., Devinney, T.M., Pillutla, M.M.: A formal model of trust based on outcomes. The Academy of Management Review 23(3), 459–472 (1998)
4. Castelfranchi, C., Falcone, R.: Trust and control: a dialectic link. Applied Artificial Intelligence 14, 799–823 (2000)
5. Clark, H.H., Schaefer, E.F.: Contributing to discourse. Cognitive Science (1989)
6. Crocquesel., C., Legras., F., Coppin, G.: Dynamic trust evaluation in supervisory control. In: HUMOUS 2010: Humans Operating Unmanned Systems, Toulouse France (2010)
7. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. International Journal of Human-Computer Studies 58(6), 697–718 (2003)
8. Jian, J.-Y., Bisantz, A.M., Drury, C.G.: Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics 4(1), 53–71 (2000)
9. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human Factors 46, 50–80 (2004)
10. Luhmann, N.: Trust and Power. Wiley, Chichester (1979)
11. Muir, B.M.: Trust in automation: Part i. theorical issues in the study of trust and human intervention in automated systems. Ergonomics 37(11), 1905–1922 (1994)
12. Muir, B.M., Moray, N.: Trust in automation: Part ii. experimental studies ofr trust and human intervention in a process control simulation. Ergonomics 39(3), 429–460 (1996)
13. Parasuraman, R.: Humans and automation: Use, misuse, disuse, abuse. Human Factors 39(2), 230–253 (1997)
14. Rempel, J.K., Holmes, J.G., Zanna, M.P.: Trust in close relationships. Journal of Personality and Social Psychology 49(1), 95–112 (1985)
15. Sheridan, T.B.: Telerobotics, Automation, and Human Supervisory Control. MIT Press, Cambridge (1992)
16. Sutcliffe, A.: Trust: From cognition to conceptual models and design. In: Martinez, F.H., Pohl, K. (eds.) CAiSE 2006. LNCS, vol. 4001, pp. 3–17. Springer, Heidelberg (2006)
17. Traum, D.R.: A Computational Theory of Grounding in Natural Language Conversation. PhD thesis, University of Rochester (1994)

A Testbed for Exploring Human-Robot Interaction with Unmanned Aerial and Ground Vehicles

Jaime H. Flores[1], Glenn A. Martin[1], and Paula J. Durlach[2]

[1] University of Central Florida, Institute for Simulation and Training, 3100 Technology Parkway, Orlando, FL 32826
[2] U.S. Army Research Institute for the Behavioral and Social Sciences, 12350 Research Parkway, Orlando, FL 32826
{jflores,martin}@ist.ucf.edu, Paula.Durlach@us.army.mil

Abstract. Over the last twenty years, the emerging roles of unmanned aerial/ ground vehicles in the U.S. military presented a number of different research opportunities in usability and training, ranging from robotic control interfaces to human-robot team collaboration. In this paper we present a testbed that we developed as a flexible software platform to explore a variety of training and coordination issues with UXVs for military application.

Keywords: Interface Usability, Unmanned Vehicles, Team Collaboration.

1 Introduction

Effective operation and control of UXVs (unmanned aerial/ground vehicles) is a major usability question, as the U.S. military is heavily reliant not only on technological superiority but also superior training in the use of that technology. [1]

In order to effectively study the impact of the UXV user interface on performance and collaboration of an operator with a commander, a testbed should provide interchangeable input mechanisms for controlling a vehicle, multiple types of sensor feedback regarding the vehicle state, and control over the presentation of vehicle data. A flexible feedback system should also be included to present experimenter feedback to the operator.

Beyond single-operator interface questions, team training and collaboration aspects are also investigated due to the ubiquity of squad-level missions in the military. This centers on the interaction of an operator-commander pair of participants, allowing the pilot to control one or more UXVs while a co-federate provides mission plans with detailed routes, areas of interest, and specific points over a geographic area. Communication capabilities between the participants should also be taken into consideration as part of a testbed, since the participants may need to work in a distributed setting.

As part of research opportunities with the Army Research Institute for Behavioral and Social Sciences, the testbed was developed for the purpose of studying the effectiveness of current and experimental robotic control interfaces and their use within joint human-robot teams. The rest of this paper presents an in-depth discussion

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 514–521, 2011.

of our testbed for this research, with emphasis on the versatility in the design and future expansions on this work.

2 Platform

Our platform consists of three software applications that have been developed over the last six years. Each application is designed to provide a specific, yet flexible feature set geared toward research in UXV usability and training. SimOCU (Simulated Operator Control Unit) is the main simulation application presented as the operation control interface for one or more UXVs. This application includes interchangeable models for controlling the motion of the unmanned vehicle, as well as a camera model based on a pan/tilt/zoom metaphor. C2Node (Command and Control Node) provides an interface for a commander participant to work along with the UXV operator, providing mission planning and support both before and during a mission. AFS (Automated Feedback System) adds a fully configurable rule-based feedback generation system to the platform, using an Action-Response interface metaphor.

2.1 Simulated Operator Control Unit

SimOCU is the core application for simulation and testing of different user interface and control configurations for an unmanned vehicle. It is designed to either work in standalone or as part of a distributed simulation, with multiple control and motion models, and configurable frame-based layout to allow for wide range of control in experimentation.

Virtual Environment. The simulation space is a virtual environment created using the Virtual Environment Simulation Sandbox (VESS) [2] and supports up to three different UXVs per simulation. Each vehicle can be configured to simulate a variety of sensors, notifications, and behaviors, including number of cameras, out-of range warnings, flight status sensors, and transmission quality. Warnings and flight status can be configured to specific distances or commands relative to the UXV base station, and the transmission latency can be set to a specific delay to simulate real-world communication variables. The environment can further be configured for a variety of real world areas and scenarios through loaded graphical models, as well as control over time of day, fog/visibility, and some basic weather patterns.

The camera model is based on the pan/tilt/zoom metaphor, and can be configured to allow a range of movement (or disabled entirely) on each axis per UXV. Up to three cameras are supported per vehicle, and each can be positioned on the simulated model to present an accurate display of the orientation and view each camera is expected to show. The controls also allow snapshots to be taken from any camera and can be transmitted to (as well as receive from) the C2Node application.

SimOCU is also built on the GEMINI library [3], making this application both DIS and HLA compatible. As such, this application can be used as part of a larger distributed simulation with entities capable of being displayed in the SimOCU application.

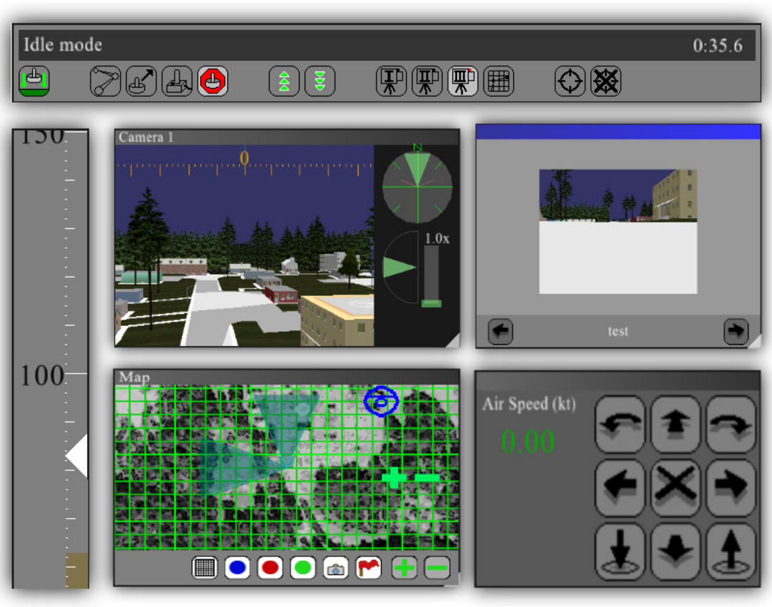


Fig. 1. Available feedback frames for UXV operation in SimOCU

Multiple Control and Motion Models. SimOCU can support different physical interfaces for controlling the vehicles, as well as different motion models for each vehicle. In particular, a participant can control any UXV via a traditional computer keyboard and mouse, or via a USB game controller (similar to those included with video game consoles).

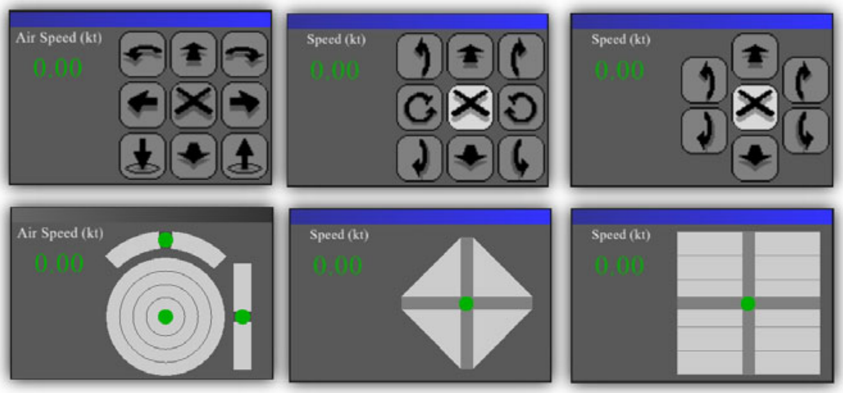


Fig. 2. Manual control interfaces in SimOCU

Motion models supported for UXV control include a continuous control scheme (where a user can control the exact speed of the vehicle on multiple axes) and a discrete control scheme (where a user can issues movement commands for a single axis). Figure 2 shows the manual control interfaces for these models currently in SimOCU for aerial, wheeled and tracked UXVs respectively, though the application is designed to easily expand support to other motion models (as well as physical control interfaces).

Beyond manual control, the simulation also supports automated navigation of the UXV through a waypoint-based spline path system. SimOCU includes an editor for creating these automated routes, as well as having assigning specific actions at each waypoint, such as, patrolling an area or taking a camera snapshot.

Configurable Frame Layout. The UXV interface is based on a number of distinct frames (shown in Fig. 1) displaying information received from the simulated vehicle. This model allows for the most flexibility in interface control as these frame elements can be moved, reconfigured, or hidden on an individual basis to allow for precise control of the size, position, and information shown in each one. The layout of these frames can be saved and reloaded on demand for any given simulation.

2.2 Command and Control Node

C2Node was designed as a complement to SimOCU by providing an interface for a commander to collaborate with the UXV operator. The application can display camera and position data for the unmanned vehicles and provides tools for creating mission plans that can be sent to UXV operator on demand, as well as text and voice communication.

Streaming UXV Data. C2Node (shown in Fig. 3) can receive UXV camera and position data transmitted through a network as it also supports DIS and HLA protocols through GEMINI [3]. It can handle data streams from any active UXV from a single SimOCU at a time. It can also send images to the operator as well as receive camera snapshots from any of the UXVs. A transmission delay can also be specified for receiving all data from the operator to simulate real world communication issues, as well as simulate loss of camera feed should the UXV move out of range from the base station, as specified in the SimOCU simulation.

Mission Planning. C2Node includes the capability to create, edit, and send a mission plan to an UXV operator. A mission plan consists of a number of markers that appear on the Map Frame to indicate a point/area of interest or to define a route the operator should take with the vehicle. This serves as a complete and effective way to communicate instructions and goals to the UXV operator beyond simple text or voice messages. Both text and voice channels however are supported for more traditional communication.

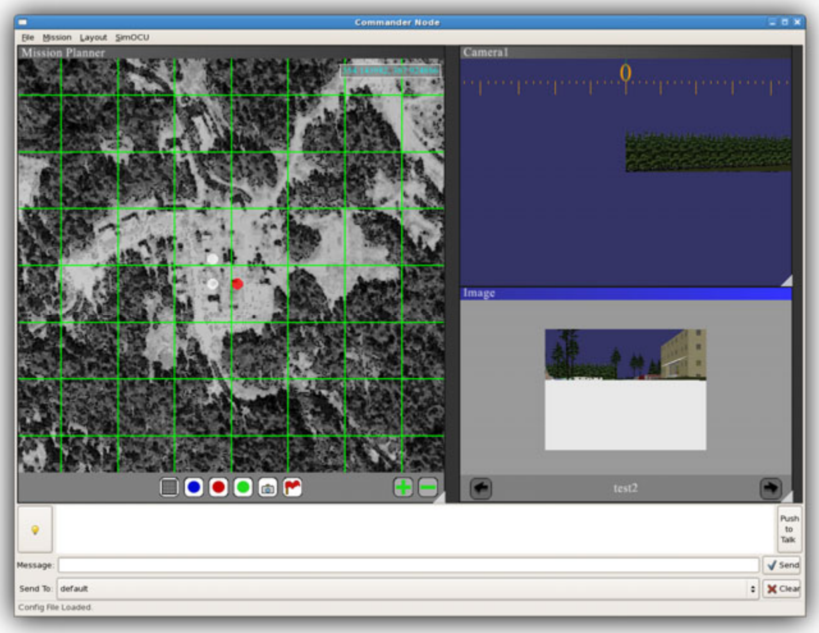


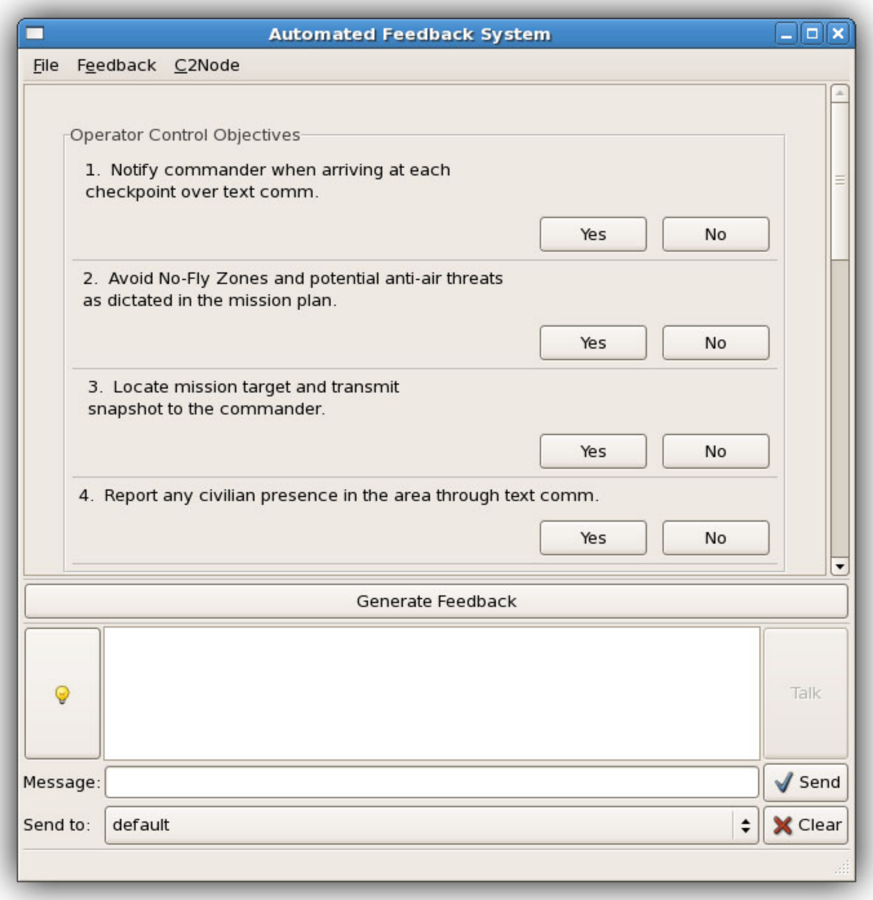
Fig. 3. The main interface in C2Node

2.3 Automated Feedback System

AFS is a rule-based feedback system that, once configured, can generate text feedback based on a user-specified set of questions used to generate raw scores. Feedback can then be transmitted to a C2Node or SimOCU user over a network.

Rule System. The generation of feedback in AFS is based on a graded Objective-Action-Response system. An Action is a collection of Responses uniquely tied to that action, and an Objective is a collection of Actions. Each level (Objective, Action, Response) can be graded separately to provide a raw score. From the various scores, Feedback is conditioned through user-specified Rules that define what kind of Feedback is generated. Objectives, Actions, Responses, Feedback, and Rules are entirely configurable, and can be saved and reloaded on demand into AFS (Fig. 4 shows an example).

Feedback Generation. Feedback is generated by the experimenter at the end of a mission and transmitted to a C2Node or SimOCU user. Along with the feedback, AFS also supports text communication to both SimOCU and C2Node. When feedback is generated, each set is normally generated completely independent of previous missions and scores. The Rule system, however, includes comparison options that allow current feedback scores to be compared to previous mission scores for more targeted and precise feedback.



Operator Control Objectives

1. Notify commander when arriving at each checkpoint over text comm.

2. Avoid No-Fly Zones and potential anti-air threats as dictated in the mission plan.

3. Locate mission target and transmit snapshot to the commander.

4. Report any civilian presence in the area through text comm.

Message:

Send to: default

Fig. 4. The main interface in AFS

3 Research and Usage

Development of our experimental testbed has always been research-driven through work with the Army Research Institute for the Behavioral and Social Sciences (ARI). Completed studies employed the testbed to examine methods and metrics for operator training, and aspects of UXV interface usability [1] [5] [6].

Initial work focused almost entirely on interface questions regarding the intuitiveness of UXV operator control unit [1]. A number of different elements including physical control (mouse vs. game controller), motion model (discrete vs. continuous), and sensor arrangement (one combined camera display vs. two simultaneous camera displays) were compared to determine their affect of each on training and performance. The findings suggested that given certain patterns that arose in the data, a heuristic could be used for designing an effective UXV control interface Subsequent work established how different interface elements affected different aspects of performance during operator training [5] [6].

Future work will use the AFS system to examine the relative training effectiveness of providing feedback adapted to the trainee's skill level and team training issues.

In order to provide a setting to examine communication and coordination issues between UXV operator and commander, a separate virtual environment simulation for running dismounted infantry missions was added. This allowed for a collective exercise with a UXV and multiple friendly, hostile, and civilian entities, which provided the commander a mission context in which to employ the UXV. The initial findings support the potential for effective training using the UXV while on a mission [4], and after some further development on the platforms, would allow more complete exploration of team coordination issues and the training which would address them. Additional questions are now also being explored surrounding the factors involved in effective human-UXV teamwork.

4 Limitations

The testbed presented here was designed with some flexibility in mind for future expansion. SimOCU in particular was developed such that the additional UXV platforms, interface elements, control schemes or simulation elements can either be done by the researcher in the application configuration, or would only require a short development cycle.

Despite this intent, the applications still have some limitations. This has been more recently apparent in AFS. While generating feedback is automated from the marked responses, it cannot automatically detect actions in the simulation that might trigger a particular kind of feedback, which is why a researcher must currently enter all the information manually. This is due to lack of infrastructure and established protocol for detecting such specific actions as taking a snapshot of a particular target, or identifying a specific object. This limitation, however, is not insurmountable.

Limitations in C2Node and SimOCU have largely been related to integration with other simulation systems. Integrating Research Network Inc.'s GDIS2 military training simulation (described in [4]) presented a number of entity synchronization issues within each application. This was in large part due to the differences in the terrain models used by each application. GDIS2 in particular uses a proprietary data format for terrain models, which prevented the model's use with SimOCU and C2Node. This meant that a custom terrain model for SimOCU and C2Node was required. Differences in these models caused problems in matching camera views and entity positions from the UXV simulation with those views seen in GDIS2. Circumventing these issues for [4] required careful adjustments to the simulation and use of the terrain models to match all the necessary views and entities. Recent developments with VESS [2] should now allow the GDIS2 terrain models to be used with SimOCU and C2Node, thus eliminating most of these problems in future research.

References

1. Durlach, P.J., Neumann, J.L., Billings, D.R.: Training to Operate a Simulated Micro-Unmanned Aerial Vehicle with Continuous or Discrete Manual Control. Technical Report 1229. U.S. Army Research Institute for the Behavioral and Social Sciences (2008)
2. Daly, J., Kline, B., Martin, G.A.: VESS: A Testbed for Virtual Reality Research and Application Development. In: Proceedings, IEEE Virtual Reality 2002, pp. 289–291. IEEE Press, Los Alamitos (2002)
3. Martin, G.A., Daly, J., Thurston, C.: Interaction within Multimodal Environments in a Collaborative Setting. In: First International Conference on Virtual Reality (2005)
4. Durlach, P.J., Priest, H., Martin, G.A., Saffold, J.: Developing Collective Training for Small Unmanned Aerial Systems Employment. In: Proceedings, MODSIM World Conference, pp. 235–240 (2009)
5. Billings, D.R., Durlach, P.J.: How Input Device Characteristics Contribute to Performance During Training to Operate a Simulated Micro-Unmanned Aerial Vehicle. Technical Report 1273. U. S. Army Research Institute for the Behavioral and Social Sciences (2010)
6. Billings, D.R., Durlach, P.J.: Effects of Input Device and Latency on Performance While Training to Pilot a Simulated Micro-Unmanned Aerial Vehicle. Technical Report 1234. U. S. Army Research Institute for the Behavioral and Social Sciences (2008)

Technological and Usability-Based Aspects of Distributed After Action Review in a Game-Based Training Setting

Matthew Fontaine, Glenn A. Martin, Jason Daly, and Casey Thurston

Institute for Simulation and Training, University of Central Florida,
3100 Technology Parkway, Orlando, FL 32826, USA
{mfontain,martin,jdaly,cthursto}@ist.ucf.edu

Abstract. After action review (AAR) in the distributed setting provides for some unique problems. Some of these problems include remote facilitation of an after action review, keeping a lightweight infrastructure that can handle large amounts of throughput and allowing for different AAR sessions to be run simultaneously. This paper proposes a method for developing a facilitative infrastructure in the AAR setting while providing a solution that allows for syncing of multiple AAR software to one review session.

Keywords: After Action Review, AAR, Simulation, Training, Software Infrastructures, Client-Server.

1 Introduction

Game-based training is increasingly seen by the U.S. military as a method to leverage the achievements of the video game industry to provide improved training at reduced cost. However, a number of technological problems exist regardless of whether game-based or older, legacy-based systems are used. In fact, solving some of these problems has become even more critical as video games themselves address other problems. In addition, a number of usability questions arise when considering a distributed setting.

Video games for training, particularly those already supporting multiplayer capabilities, provide an easy-to-arrange platform for distributed team training. Players can simply connect to a server, join an exercise and participate as one member of a team. One element missing throughout most of these games, however, is a satisfactory after action review (AAR) capability. AARs provide an opportunity for the trainees to review what happened during the exercise and identify ways to improve future performance. This is true regardless of whether the game is used purely locally or distributed across the Internet (with trainees at multiple sites).

We have been pursuing work in addressing the problem of providing an AAR in a distributed game-based training setting. While a traditional AAR system, providing a review in an auditorium setting where all trainees are physically together, presents its own set of challenges, additional challenges are presented when that system needs to provide a review session across multiple sites.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 522–529, 2011.

2 DIVAARS

Our testbed consists of nine trainee stations and one instructor station within our laboratory. In addition, we collaborate with other sites including Embry-Riddle Aeronautical University, Old Dominion University and the U.S. Military Academy at West Point. Each station can run either the Game-Distributed Interactive Simulation (G-DIS) system from Research Networks, Inc., the On-Line Interactive Virtual Environment (OLIVE) system from SAIC, Inc., or the Virtual BattleSpace 2: Army (VBS2: Army) system from Bohemia Interactive, Inc.

During an exercise using any of these systems, we use our previously-developed AAR system, the Distributed Infantry Virtual After Action Review System (DIVAARS) [1]. DIVAARS records all actions taken and verbal communication made by all participants and presents a stealth view for the AAR Facilitator to follow the exercise. Tags (bookmarks) can be placed in time to provide an easy mechanism for jumping to specific events during the review session. DIVAARS can be used in an auditorium setting for local AAR sessions and will playback each segment, providing a visual view of the exercise from any viewpoint and clearly depicting all actions taken by the trainees. All verbal communication is replayed as well. DIVAARS has been well received by Soldiers in representing what happened and supporting a discussion of why it happened and how to improve in future exercises and missions [2].

3 AAR in a Distributed Setting

When performing an after action review in a game based training setting, there are a number of issues that arise. First and foremost, the AAR host has no method of communication with the trainees at remote sites. The AAR facilitator must be able to communicate with each remote site and verify that everybody is ready to start the review and that everybody has joined the AAR session. DIVAARS originally provided visual and auditory playback of the exercise at a single site. In order to provide the same experience across multiple sites, additional software that can perform the same kind of functionality at the remote sites must be provided. This software must be connected with DIVAARS in some way so that the AAR facilitator can still conduct the AAR session as before. Since most game-based training systems simply use the Internet for communication, DIVAARS and the new remote stations would likely also be required to use the Internet. Third, since the trainees can no longer see the AAR facilitator, actions as simple as pointing to the DIVAARS screen to highlight a particular element are no longer possible. While not absolutely necessary, additional tools to replace these simple gestures may make the AAR session more productive.

With the addition of distributed training, it becomes possible to conduct larger training exercises. With this capability, the need for multiple simultaneous AAR sessions could potentially arise. Each of the smaller units comprising the larger overall training exercise could have different goals and training objectives between them, and having the capability for multiple, separate, and simultaneous AAR session may be beneficial to the individual unit leaders.

An additional requirement for a distributed AAR system is overall robustness and fault tolerance. A software problem with the AAR system at a single site can typically be resolved by simply restarting the software. This is a much more complex and time-consuming issue to resolve in a distributed situation. Thus, the distributed AAR system must be able to deal with any unexpected errors and exceptions with as little user interaction as possible. It also must perform well enough to handle the inherent latency, jitter and communications issues typical of Internet connections without introducing any significant delays of its own. The users at remote sites are likely to be novices when dealing with AAR software, so the details of creating the remote connection and handling any communications errors must be transparent to the user.

Finally, the distributed AAR system must be extensible enough to be able to add any additional software that might be needed for the review session to the system. For example, if a voice communication solution were needed to support the AAR session, the distributed AAR system should be capable of incorporating such a system without significant changes to the underlying software. While DIVAARS addresses its goals very well, it did not support performing a distributed AAR in its previous form. In order to support distributed AAR sessions, we needed to add new features to DIVAARS to address the issues detailed above.

4 DART

Our first approach to addressing the technological issues of providing for a distributed AAR focused on using our AAR software engine, SOCRATES, to create a new application [3]. SOCRATES is a fully modular, plug-in based architecture that utilizes a message passing mechanism to communicate between the plug-ins. The Distributed AAR Remote Tool (DART) is essentially an additional DIVAARS station with a simplified user interface, intended for use by the trainees. It included a plug-in that would directly connect to the master DIVAARS station that the AAR Facilitator used, and listened to the same messages that were used to run the master station. The result was effectively a mirror-image of the main DIVAARS station. As the facilitator moved the main view, the same view was drawn on any connected DART stations. When the facilitator jumped to a particular event in DIVAARS, the DART stations also jumped to the same event. We also provided a telestration tool to allow the instructor to draw on the screen to highlight a particular event or item (replacing the simple gesture that was possible in the auditorium setting). Additionally, we added a simple voice intercom system, allowing the trainees and facilitator to communicate verbally. The result worked reasonably well in the controlled conditions of a lab and local-area network. In practice, however, this implementation encountered a number of problems.

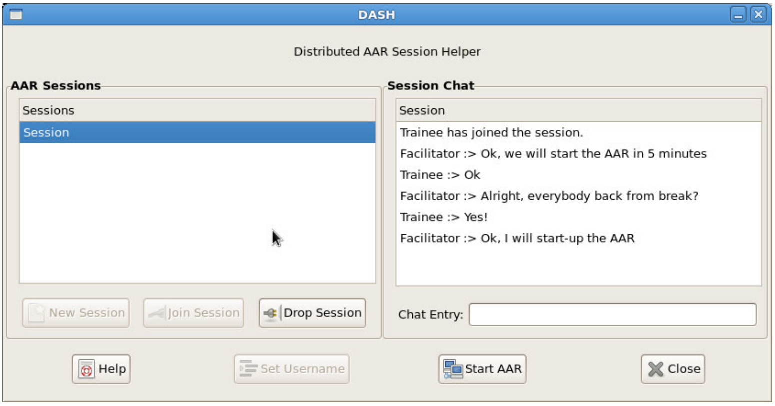
First, the master DIVAARS station must handle the normal processing of the review session for the Facilitator, as well as communicating with every DART client attached to the review session. Our first implementation had the DIVAARS station transmit data to each client in series. If one client had a faulty or slow network connection, DIVAARS would pause while it waited for that client to catch up. Placing the processing of remote client connections on a separate thread only helped on the master station (each client still must be served in series). We also considered a kind of round-robin scheduling scheme, where messages were queued for the problematic clients until they could catch up. However, this would have required large amounts of additional memory in a system that already uses a significant amount. Another problem with the simple, direct-connection model is not being able to run multiple after action review sessions simultaneously. If a second session was desired, an additional computer would have to host the session and a different set of connection information (IP address, port, etc.) would need to be provided to each trainee. This can potentially be confusing if the trainees aren't sure as to which session they need to connect.

A final problem that could occur was in network connection coordination. The TCP protocol includes certain provisions needed to provide reliable network communication. These include connection handshaking and connection time-out features. If one side of a connection drops unexpectedly, the connection is kept open for a defined time period in the case that the interruption is temporary and latent packets might still arrive. This can cause problems in coordinating the restart of AAR applications after a crash. For example, if the DIVAARS station needed to be restarted for any reason, each and every DART station would also have to be closed and restarted; otherwise, left-over TCP connections would prevent new connections from properly being connected until the old ones timed out, which can take several minutes. On the surface, this sounds like a minor issue, but in various trials, it was one of the most noticeable problems with the original distributed system. This was compounded by the fact that after a given station crashed, no instructions could be given to that station (recall that all communication was handled through the software itself), so the trainee did not know whether to wait, to restart the software, or to do something else.

These problems prompted a redesign of the distributed AAR system. The new design incorporated new features to address each of the problems.

5 A New Approach

The Distributed AAR Session Hub (DASH) was developed as a means of addressing some of the issues when running distributed after action reviews. It enables a smooth transition between the training scenario and the after action review. To handle the issue of pre-AAR communication Wiese suggests at the minimum audio chat and ideally voice communication be supported for communication for after action review. [4] When writing DASH it was decided that a text messaging style chat room to be more ideal. If the instructor were to ask if everyone were ready and some trainees were not ready they could miss the message. Using a text based message system the trainees can scroll throughout the conversation looking up discussions they missed when they were not ready. During the after action review DIVAARS already supports audio chat. In previous exercises, each trainee was required to exit the game-based simulated training environment, manually run the DART client, and connect to DIVAARS using controls in the DART client itself. DASH acts as a pre-AAR "lobby" much like pre-game lobbies for multiplayer online video games. Instead of providing a hostname or IP address, the user simply selects which session to join by name. Once inside the AAR session, the Facilitator and trainees can chat back and forth to confirm that everyone is ready to start the AAR. The chat can also serve as a way for trainees to communicate if the AAR session goes down or if the trainees have any other problems that cannot be addressed through the AAR review software itself.



Trainee has joined the session.
Facilitator :> Ok, we will start the AAR in 5 minutes
Trainee :> Ok
Facilitator :> Alright, everybody back from break?
Trainee :> Yes!
Facilitator :> Ok, I will start-up the AAR

Fig. 1. DASH chat

A screen shot of the DASH client is shown in Figure 1. The facilitator has created a new session and the trainee has joined it. The facilitator has asked if everyone is ready and just needs to press the Start AAR button to run DIVAARS. DASH runs DIVAARS on the Facilitator's station and DART on each client station. DASH automatically synchronizes all stations to the chosen AAR session.

In addition to the lobby and coordination aspects to DASH, it also provides a server capability of the AAR session itself. Rather than depending on DIVAARS to perform the server needs, DIVAARS now simply becomes another client, passing all of its messages to the DASH server. The DASH server then relays the messages to each client. This allows problematic clients and bad connections to be addressed without affecting the performance of the overall AAR session itself.

5.1 Implementation

In implementing the server side of DASH, we wanted to address several requirements. DIVAARS requires a large amount of data throughput, so we needed to keep the server as lightweight as possible. DASH also introduces the concept of an AAR "session" in order to abstract the low-level details of network connections away from the users. DASH needed to maintain this "session" concept regardless of the type of AAR software that utilizes it (we wrote DASH with DIVAARS in mind, but its utility is not limited to only DIVAARS). The server must be able to keep track of a number of AAR sessions that may be running simultaneously. The DASH server is composed of a number of modules, each providing its own service and running in its own thread of execution. Currently, there are three modules within the DASH server. The chat server connects the "lobby" of each DASH client with the other clients in the session, and allows the text chatting described previously. The AAR server connects the DIVAARS and DART clients together and passes the AAR messages between them. A third module, the client server, is responsible for tracking client connections, the services they are using, and the sessions to which they subscribe. This is shown in Figure 2.

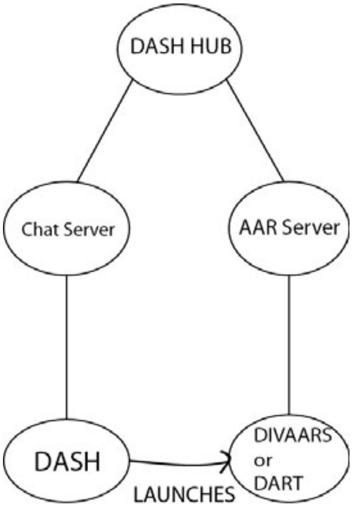


Fig. 2. DASH

DASH sessions are created by the instructor on his or her DASH client. On the server side, the chat server is responsible for creating the sessions requested by the instructor. This notion of which session each client belongs to must be shared with each of the other servers. To accomplish this, the chat server informs the client server whenever a given client joins a session, and the client server maintains a list of clients for each session. When another server needs to access client information, it must obtain a client list from the client server.

The client lists are stored on what can be thought of as a virtual bookshelf. When a server needs a particular client list, it must first check out the client list. The server is then given mutually exclusive access to the client list. A server can only be given access to one client list at a time. This is done in order to prevent deadlock between the servers.

Other options were considered for synchronization of client data, including message passing between servers and virtualizing the client lists for use between different servers at once. We decided that the virtual bookshelf model had the lowest overhead and best met our goal for maintaining a lightweight server.

Every time a client connects on the AAR server the client list of the chat server is checked out. The AAR server can then synchronize the client's session to the client in the chat server's client list. The synchronizing of clients is done simply by the fully-qualified hostname. This all happens transparently to the users. This is shown in Figure 3.

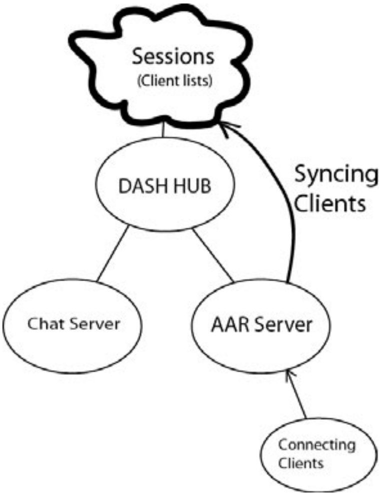


Fig. 3. Syncing

To make the data passing on the AAR server as fast as possible, we compress the data before sending it to the server using the well-known zlib compression library. Because of this, the data is never read by the DASH server. In order to determine to which client or clients each incoming message must be sent, DASH simply checks the client list containing the sending client, and relays the message to all other clients in the list.

The design of the client server provides extensibility to the overall DASH server hub. Any new servers can be developed independently and easily added to the hub. The new server simply adds its client data to the client server, so other servers can access it. For example if we wanted to add a server to support video chat software like Wiese suggests but as a separate application from the review software we can easily connect it to the session. [4]

5.2 Limitations

The virtual bookshelf model for client tracking does have limitations. In order to eliminate deadlock and reduce overhead, the DASH server has to limit its access to one client list at a time. This eliminates the possibility of concurrent communication between servers. Also, the model only allows for synchronizing client information between DASH servers. There is no provision for other kinds of messages. For example, it may be desirable for a DASH client to know if and when the AAR program crashes. This would require the DASH AAR server to communicate with the chat server that a client crashed.

Another limitation of the current implementation is that we are currently unable to run each session on each server in its own thread. Although the message handling in DASH server is a producer-consumer model, it does not handle this on a per-session basis between threads. If one AAR session begins to bog down, the effect could possibly be felt by another session that is also running on the same server.

6 Potential Problems

The largest amount of throughput when using DIVAARS occurs when the AAR facilitator fast-forwards or rewinds the scenario. DIVAARS uses state information of units to update entity position and orientation in the environment. In a naïve approach, this state information is passed would be passed to the connecting DART client for each change in state. This would be fine when the instructor moves through the scenario at a normal pace since the entity information is just passed at a normal rate to the connecting clients. However, in a fast-forward (or rewind), state information is processed more quickly (since time is passing at a faster rate). The system essentially would have to "catch up" in the scenario to the new position of all the entities. This would cause a problem because the large factor increase in the number of entity updates that need to be sent out. The same number of entity updates would be sent through the system as it would when playing the scenario normally. To avoid this issue we use dead reckoning within DIVAARS (much like a typical simulator might) that reduces the amount of information that is sent out to its distributed AAR clients.

7 Conclusions and Future Work

Our work has created a system that allows for easy facilitation of distributed after action reviews. The server side of DASH also provides for a unique infrastructure for hosting many different kinds of software that need to be linked to the same session as the after action review.

In the future we would like to further improve the computational capabilities of the DASH server. In addition, we would like to evaluate how well the coordination works and if trainees find any confusion in its operation. We also plan to investigate the Distributed Debrief Control Protocol (DDCP) that is currently being developed by the Simulation Interoperability Standards Organization (SISO).

Acknowledgements. The authors would like to acknowledge the support of the Army Research Institute (ARI) and the Office of Naval Research (ONR) who helped partially fund this work.

References

1. Knerr, B., Lampton, D., Martin, G., Washburn, D., Cope, D.: Developing an After Action Review System for Virtual Dismounted Infantry Simulations. In: I/ITSEC (2002)
2. Lampton, D., Bliss, J., Martin, G.: Performance Measurement and Training Feedback in a Military Collaborative Virtual Environment. In: HCI International (2005)
3. Martin, G., Daly, J., Thurston, C.: An After Action Review Engine for Training in Multiple Areas. In: HCI International (2011)
4. Weise, E., Freeman, J., Salter, W., Stelzer, E., Jackson, C.: Distributed After Action Review for Simulation-Based Training, Human Factors in Simulation and Training, ch.15. CRC Press, Boca Raton (2009)

Authority Sharing in Mixed Initiative Control of Multiple Uninhabited Aerial Vehicles

Rui Gonçalves[1], Sérgio Ferreira[1], José Pinto[1], João Sousa[2], and Gil Gonçalves[1]

[1] Department of Informatics Engineering
[2] Department of Electrical and Computer Engineering
School of Engineering, Porto University (FEUP)
Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
{rjpg,ei05092,zepinto,jtasso,gil}@fe.up.pt

Abstract. In this paper we discuss a conceptual framework that supports operational scenarios with multiple UAVs and operators. These UAVs possess different levels of autonomy while the operators have variable skill sets. The scenarios themselves encompass different missions, with different phases (requiring different levels of attention from the operator) and with the occurrence of various exogenous events. This framework was employed in the development of a Command and Control (C2) application which is capable of operator advisement, self adaptation, and automatic task distribution among operators and UAVs, depending on mission objectives, phase and occurrences. This C2 application enables a clear overview of the remote environment by placing the operator closer to the control loop, whether it is at an abstract or low level of control. Consequently there is an improvement of task redistribution and situation awareness, as well as reduction of workload.

Keywords: Operator, UAV, Interoperability, Autonomy Levels, Command and Control, UAS, Situation Awareness, Workload.

1 Introduction

The last decades have witnessed unprecedented technological developments in computing, communications, navigation, control, composite materials and power systems, which have led to the design and deployment of the first generations of unmanned aerial vehicles (UAV) and unmanned aerial systems (UAS). These vehicles have already seen action in many scenarios and proved their value.

As the operational capacity of UAS continues to grow, these systems can include multiple UAVs operating as a team, furthermore solidifying their employment in military and civilian scenarios. With the aid of these systems it is possible to remove the human element from “dirty, dull, and dangerous” situations and relocate it to a less operational and more supervisory role. However, with the rise of their operational capacity so rose the complexity of tasks they could perform.

At the Underwater Systems and Technology Laboratory (LSTS) [1] we have been designing, building and operating a significant number of heterogeneous unmanned

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 530–539, 2011.

vehicles. These include Remotely Operated Vehicles (ROV) [2], Autonomous Underwater Vehicles (AUV) [3, 5, 6], and Autonomous Surface Vehicles (ASV) [4].

We have been also developing UAVs [7] as a result of our collaboration with the Portuguese Air Force Academy.

Throughout this paper we describe a conceptual Framework for optimal inclusion of the operator in the control loop and the application of its concepts in a C2 software interface. The objective is to distribute and reduce the workload of a decentralized team of operators controlling multiple UAVs. To achieve this goal we intend to advise operator's actions and reconfigure C2's layout using an automated methodology. The combination of the different mission intervenient entities (UAVs, Plan State, Operators, Consoles Profiles, and Mission Workload) can be used to inform the operator about the ideal operation console layout to be used and the ideal workload for each operator. The properties of these mission entities can change, during the mission execution, making this a dynamic process. The operator can have different levels of situation awareness, at different stages of the mission. The system will help operators to dynamically configure an optimal view of the mission state from a set of predefined console layout profiles.

In section 2 we introduce the C2 framework inner workings and its operation method. In section 3 we give a complete overview of the LOA framework and it's execution principles. In section 4 we present an example of the LOA framework at work and in section 5 we introduce two examples of operation consoles, to be used in conjunction with de C2 framework.

2 Networked Vehicle Systems and Supervisory Control

Unmanned vehicle systems are currently being employed in the field for very distinct purposes. For instance, considering just individual UAVs, these can be used for precision sensing, aerial imagery, surveillance, etc. The full potential of these systems, however, requires the management of multiple networked vehicles operating as a whole, sharing their workload and knowledge about the environment.

The concepts of operation for multi-UAV teams differ from single UAVs in the sense that in the former there exist common objectives like maintaining a common knowledge database [8] and redundant execution of crucial actions [9]. Moreover, operators are required to quickly perceive the entire system state, so that they can re-organize themselves in the face of unpredicted situations. All this while taking into account the different levels of attention all the vehicles demand. In order to decrease the number of operators' necessary on a multi-UAV deployment, we use mixed-initiative interaction for controlling the network at a system-level.

2.1 Simultaneous Control of Multi-UAV Teams

In our C2 framework, UAVs can be tasked either individually by an operator or they can be tasked by a software agent that acts as an operator (Team Supervisor). The team supervisor divides work among the vehicles according to a multi-UAV mission specification and simple task-allocation algorithms. If the control over the UAV is not overridden, they carry out planned behavior until they are faced with failures, or any other unpredicted situations in which they contact the ground station and require human intervention. When operators have sufficient authority, they can cancel the current vehicle's planned behavior and replace it with other tasks or tele-operate the vehicle (control override). This may result in the cancelation of tasks that were generated by the team supervisor and thus they will be postponed for execution by another free UAV.UAVs may also actively contact the base station asking for human intervention when there is an onboard malfunction or a potential risk is detected. In this case, the C2 framework will try to allocate the vehicle supervision to a free operator or will suggest switching of coupling between vehicles and operators.

2.2 Team Supervisor

To provide system-level control of multiple vehicles, we use a software agent that holds a multi-UAV mission specification. This mission specification is currently a list of individual plans that need to be executed by UAVs. The importance of this software module is that it allows the interaction with UAV network simply by adding plans that need to be carried out. The team supervisor then captures the capabilities among existing UAVs, their current tasks and also the availability of operators. Tasks are divided among UAVs in a way that workload is shared among capable vehicles. Some tasks however also require the intervention of human operators for correct execution, so the availability of operators is taken into account by the team supervisor while tasking the network.

3 Concepts for the Framework

This section presents how we interpreted and adapted the original LOA matrix (Table 1) into our Framework for optimal inclusion of the operator in the control loop. We will describe the LOA matrix and how we intend to categorize the operator skills. Then we describe the methodology used to advise one Console Profile (CP) to the operator for a given LOA on a mission stage, combined with the Operator Skills data. The Operator Workload management is made by the Mission Team Supervisor as described in the previous section.

3.1 Levels Definition

The LOA Table [10] is based on Sheridan's 10-level of autonomy scale [11] and simplified to present only eight levels of autonomy. The lower the task is on the scale, the more authority the human operator has over the automate. The two dimensions of the matrix (Table [10]) are the eight levels (matrix rows) crossed with four functional categories (matrix columns). The second dimension presented in this matrix is the division of each task into four functional steps. These tasks present human decision-making processes as a set of OODA cycles (Observe, Orient, Decide, and Act).

Table 1. Partial LOA matrix as originally published in [10]

Level	Observe	Orient	Decide	Action
8	The computer gathers, filters, and prioritizes data without displaying any information to the human.	The computer predicts, interprets, and integrates data into a result which is not displayed to the human.	The computer performs ranking tasks. The computer performs final ranking, but does not display results to the human.	Computer executes automatically and does not allow any human interaction.
		•••		
		•••		
1	Human is the only source for gathering and monitoring (defined as filtering and prioritizing) all data.	Human is responsible for analyzing all data, making predictions and interpretation of the data.	The automate does not assist in or perform ranking tasks. Human must do it all.	Human alone can execute decision.

Table 2 is used to categorize the operator skills using the LOAs he is certified to respond, the CP (CP-Console Profile) the operator is familiarized and the number of vehicles he can handle safety at a certain LOA. With this data we can infer about the training and education level. Notice that the LOA entry table, of operator skills, has correlation whit the number of vehicles the operator can handle.

Table 2. Fields used to infer about the operatos skills in the framework

Certified Type of LOA	Certified Consoles Profiles	Number of Vehicles
Type of maneuver the operator is certified.	Set operation Consoles the operator is familiarized. By preference order. (for one LOA)	Operator fan-out of vehicles (for one LOA)

3.2 LOA Combination and Consoles Profiles

The matrix represented in Table 1 can be related with the creation of different types of console profiles. Different console profiles can be associated to different combinations of the four functional categories (OODA) - operational modes. For the presented framework we have a direct relation of LOA and CP.

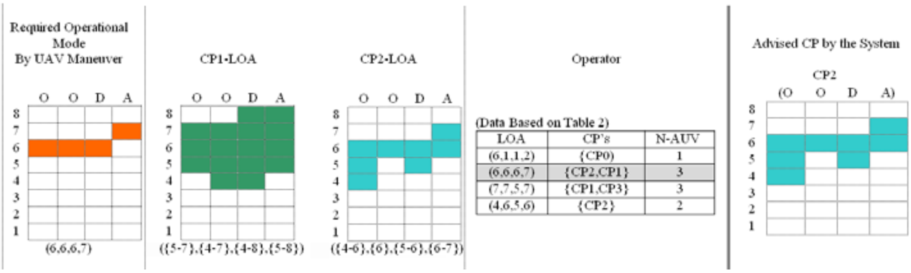
The formal representation for CP-LOA tuple is:

$$\text{CP-LOA}=(\{\text{Obs}_1\ldots\text{Obs}_n\},\{\text{Ori}_1\ldots\text{Ori}_n\},\{\ \text{Dec}_1\ldots\text{Dec}_n\},\ \{\text{Act}_1\ldots\text{Act}_n\ \})$$

For example a CP specialized for UAV "*fly-by-wire*" (direct control) operational mode would be based on the following tuple: CP-LOA= ({1}, {1}, {1}, {1}) for OODA combination. The elements on the tuple are represented as sets so we can group the OODA functional categories. This way it is possible to have one CP capable of handling different Operational Modes. Grouping some intervals of OODA levels in the LOA matrix has proved to be useful in practical application. Another example of grouping the LOA of the proposal matrix in [10] can be consulted in [12].

One high level control CP should be able to handle high LOA values for OODA. For example one representation of LOA for a console of this type can be: CP-LOA= ({5-6}, {6}, {5-6}, {6-7}). In conceptual terms we can have one very generic CP that responds to all possible combinations of LOA (CP-LOA= ({1-8}, {1-8}, {1-8}, {1-8})). Since the system is composed by several CP's, one LOA required by the UAV can have different CP's to handle interaction. This means the system will have different CP's to choose, to advise the operator to use, in a given Operational Mode of the plan state. This choice can be automated by looking to the preference order of CP's that the operator is certified.

In Fig. 2 is illustrated the processed of advising on CP to the operator. First the UAV starts a manoeuvre which requires a LOA. The system searches for the catalogued CPs in the system and operators capable of handling that LOA (manoeuvre). The listed CPs are filtered by the ones that the operator is certified. Finally the system advises the best CP to the operator. The operator is selected by the Mission Team Supervisor based on the workload of the mission operators.



(Data Based on Table 2)

LOA	CP's	N-AUV
(6,1,1,2)	{CP0)	1
(6,6,6,7)	{CP2,CP1}	3
(7,7,5,7)	{CP1,CP3}	3
(4,6,5,6)	{CP2}	2

Fig. 1. Decision process of Console Profile to advise the operator

4 Scenario Definition

To exemplify the framework's execution we will use one mission scenario where the operators have to find a target and follow it. There will be two operators and five UAVs in this example.

Current UAVs offer little adaptability in terms of automation: operators can leave the UAV do the flight by itself, following a pre-defined flight plan, or they can control it manually. For this example we will use 2 LOAs for the operators and another one of full autonomy, for handover and emergency manoeuvres. The operators LOAs are further sub-divided into a high level control LOA and low level control LOA. All three LOAs are described as follows:

- **Operational Mode 1 –** Tele-Operation or Direct Control : LOA= (3, 2, 2, 2)
- **Operational Mode 2 –** Survey : LOA= (6, 6, 7, 6)
- **Operational Mode 3 –** Full Autonomy : LOA= (8, 8, 8, 8)

Once the target is identified by the operator Operational Mode 1 will be used to follow it. When the vehicle is in "search mode" the operator sees the payload data (video) and tries to identify the targets, this is Operational Mode 2. In Operational Mode 2 the operator can also defined survey areas for each UAV. Finally Operational Mode 3 is a full autonomy mode used for the handover of UAV control logic. Its premise is that the operator must free the UAV so other operators can own it (request control of it). In this mode the system only knows about the UAV existence.

We will use two CPs (CP1 and CP2) to handle this mission example as follows:

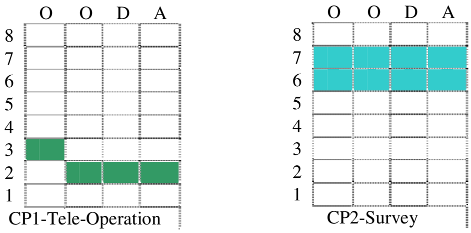


Fig. 2. Two Console Profiles used in mission

For this mission example we will have two operators with the following Skill Tables:

Table 3. Skills Table for Operator 1, based on table 2 – This operator can handle 3 UAVs in high level control and 1 UAV in low level control

Certified Type of LOA	Certified Consoles Profiles	Number of Vehicles
(3,2,2,2)	{CP1}	1
(6,6,7,6)	{CP2}	3

Table 4. Skills Table for Operator 2, based on table 2 – This operator can handle 4 UAVs in high level control

Certified Type of LOA	Certified Consoles Profiles	Number of Vehicles
(6,6,7,6)	{CP2}	4

Fig. 3 illustrates the 5 most important steps taken when one of the operators find the target. Initially all the UAVs are in survey mode – mode 2 of our LOA definition – and both of the operators are using CP2 to operate the UAVs: defining survey areas and looking at the payload data (video). In step 2 Operator 1 finds the target, which must requires on UAV to enter mode 1.This leads to a workload overload for Operator 1 that must be solved by the mission Team Supervisor. The only operator capable of handling UAVs in mode 1 is Operator 1, as defined in table 3. Since Operator 1 is only capable of handling one UAV in mode 1 the mission supervisor will advise Operator 1 to hand-over the other 2 UAVs to Operator 2. Here starts step 3 with the handover process: Operator 1 frees the UAVs putting them in mode 3. Finally, in step 5, Operator 2, takes over these UAVs in mode 2 and Operator 1 can now follow the target. In this step the Mission Supervisor advises Operator 1 to use CP1-Tele-operation to respond mode 1 LOA, according to his skills table 3.

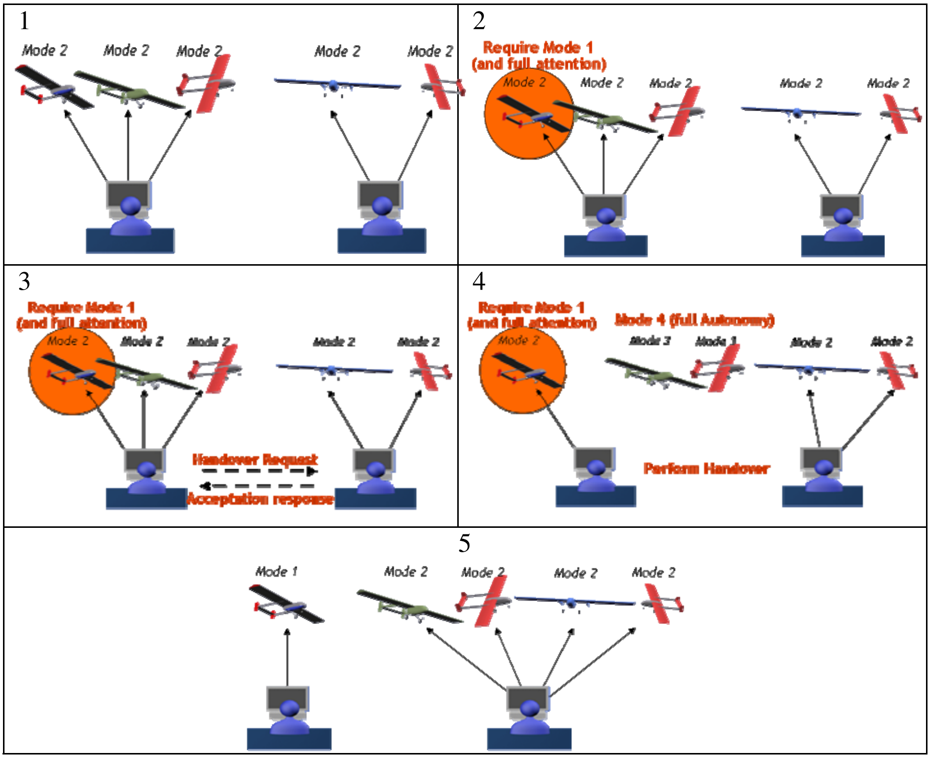


Fig. 3. Logic of operation example for mission workload distribution

5 Framework Components

As stated before, this framework was employed in an existing C2 application, Neptus, which has an underlining architecture and provides de means of creating the various consoles used in the different CP's. This section introduces Neptus and an example of such consoles.

5.1 Neptus

Neptus is a distributed C2 framework for operations with networked vehicles, systems, and human operators. Neptus supports all the phases of a mission life cycle: planning, simulation, execution, and post-mission analysis. Moreover, it allows operators to plan and supervise missions concurrently [13].

Furthermore, Neptus encompasses a console building application which facilitates the rapid creation of new operation consoles for new vehicles with new sensor suites, as well as the remodelling of old consoles for current vehicles. There are two important aspects to console configuration: visual components and event communications. The internal Neptus event communication system is based on a tree structure (following the blackboard design pattern [14]), where nodes indicate the subject of data values in leafs.

Neptus visual components can become listeners of a single variable (tree leaf) or of a defined variable domain (tree branch). Whenever a message arrives, using the IMC[15] communication protocol, that data is stored in a specific branch of the tree and listeners are then informed of the incoming network data. In a similar way, output data is sent by Neptus console components through the variable tree. The variable tree system is also used for event communication between Neptus console components.

There are two states in the Neptus generic console builder application: Editing and Operational. In Editing mode, users can then add and place components freely inside the console's main panel. Component properties can be edited to connect the panels to different systems and variables. When all components are ready, correctly placed and connected to the system variable tree, the user can switch the state of the application to the Operational mode.

5.2 Operation Consoles

Besides having the capability of dynamically creating new consoles during a specific mission, Neptus also has predefined consoles already available for the LOA switches the presented framework requires. These consoles go from standard tele-operation consoles, as seen as example 1 of Fig. 4, to supervision consoles, as seen as example 2 of Fig. 4. These consoles have different layouts depending on the central function they have. For instance a tele-operation console will typically have more detailed data about the UAV under its control, whereas a supervision console will only have a simplified view of the current UAV to allow a broader view of the whole team.

As an example of said consoles we introduce de details behind the current flight manager console used for UAV mission supervision at the LSTS.

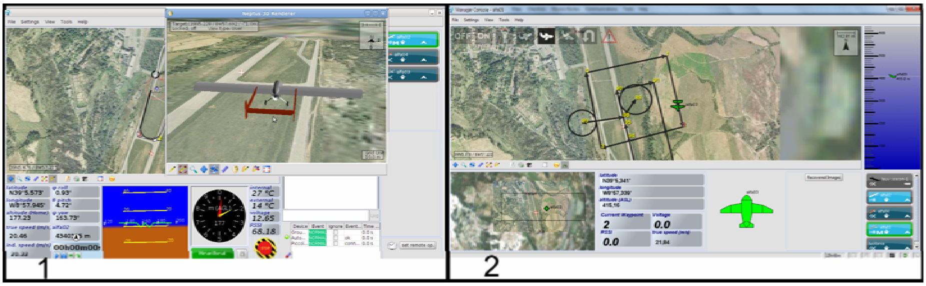


Fig. 4. Tele-Operation and a Supervisory control consoles

This console, as seen in Fig 4, was developed based on an RTS paradigm with the intent of applying the concepts, learned by these types of games, on how to efficiently control and supervise groups of units of various dimensions and with different capabilities. This approach, while not being new, has allowed the implementation of a console which supports high LOA levels CP-LOA= ({6-7},{6-7}, {6-7}, {6-7}) while, at the same time, enables the supervision of UAV teams with a low workload rating value for the operators, as can be seen in Fig. 5.

On par with workload evaluation there as has been, as well, situation awareness evaluations of these consoles in order to guarantee flight manager focus and to maximize UAV team fan-out.

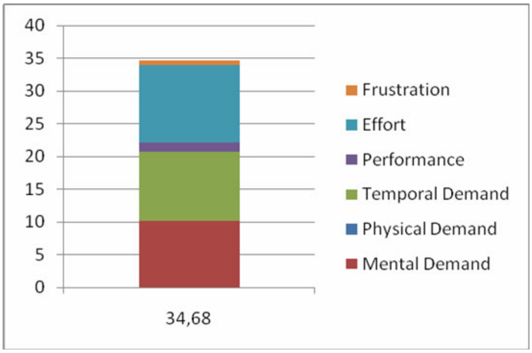


Fig. 5. Flight manager console's total workload rating, using NASA-TLX [16], in a 3 UAV scenario

6 Conclusions

Throughout this paper we referenced the growing importance of multi-UAV systems, paying special attention to the needs of the successful use of these systems.

We presented the concepts behind a framework for managing UAV task and workload allocation between various operators in a mission scenario. This framework was applied in the development of a command and control (C2) application which is capable of self adaptation, operator advisement and automatic task distribution amongst operators and UAVs depending on mission objectives, phase and occurrences. An example scenario of this framework, as well as an example of the details around one of the consoles used by the operators, was presented and discussed.

This C2 application enables a clear view and presence on the remote environment by putting the operator much closer to the control loop, whether it is high level or low level control, with the consequent improved redistribution of tasks and situational awareness. NASA Task Load Index (TLX) was used as a means to determine the adequacy of the C2 interface and functionalities. The preliminary results obtained with this framework are promising and we are confident its use will vastly improve the reliability of multi-UAV teams by augmenting their compatibility with more mission scenarios.

Acknowledgments. The authors acknowledge the contribution of AsasF of FEUP in the requirements identification and in the demonstration tests. Also acknowledge the Portuguese Air Force Academy (AFA) in the field tests and developments and design of ANTEXs X02 UAVs under the *PITVANT* project funded by the Portuguese Ministry of Defense. Neptus relies on other components of the LSTS tool chain for its correct execution. For this, we would like to thank Eduardo Marques, Paulo Dias, and Ricardo Martins mainly for their efforts in developing the IMC protocol. Rui Gonçalves and José Pinto were partially supported by FCT grants.

References

1. USTL. Underwater Systems and Technology Laboratory (February 2011), `http://whale.fe.up.pt`
2. Gomes, R., Sousa, A., Fraga, S.L., Martins, A., Borges Sousa, J., Lobo Pereira, F., et al.: A New ROV Design: Issues on Low Drag and Mechanical Symmetry. In: Oceans 2005, Europe, June 20-23 (2005)
3. Cruz, N., Matos, A., Borges de Sousa, J., Lobo Pereira, F., Estrela Silva, J., Coimbra, J., Brogueira Dias, E.: Operations with Multiple Autonomous Underwater Vehicles: The PISCIS Project. AINS
4. Ferreira, H., Martins, R., Marques, E., et al.: Swordfish: an Autonomous Surface Vehicle for Network Centric Operations. In: IEEE Oceans Europe (2007)
5. Madureira, L., Sousa, A., Sousa, J., et al.: Low Cost Autonomous Underwater Vehicles for New Concepts of Coastal Field Studies. In: CERF ICS (2009)
6. USTL. Seascout (January 2010), `http://whale.fe.up.pt/seascout`
7. Pereira, E., Bencatel, R., Correia, J., et al.: Unmanned Air Vehicles for Coastal and Environmental Research. In: CERF ICS (2009)
8. Jariyasunant, J., Pereira, E., Zennaro, M., Hedrick, K., Kirsch, C., Sengupta, R.: CSL: A Language to Specify and Re-Specify Mobile Sensor Network Behaviors. In: Proceedings of RTAS (2009)
9. Sousa, J.B., Simsek, T., Varaya, P.: Task planning and execution for UAV teams. In:Proceedings of CDC (2004)
10. Proud, R.W., Hart, J.J., Mrozinski, R.B.: Methods for determining the level of autonomy to design into a human spaceflight vehicle: A function specific approach. In: Performance Metrics for Intelligent Systems Workshop, Gaithersburg, MD (2003)
11. Sheridan, T.B.: Telerobotics, automation, and human supervisory control. MIT Press, Cambridge (1992)
12. Villaren, T., Madier, C., Legras, F., Leal, A., Kovacs, B., Coppin, G.: Towards a Method for Context-Dependent Allocation of Functions. In: HUMOUS 2010 conference Humans Operating Unmanned Systems ISAE - ONERA, Toulouse, France, April 26-27 (2010)
13. Dias, P.S., Pinto, J., Gonçalves, R., Sousa, J.B., Pereira, F.L., Gonçalves, G.: Neptus, command and control infrastructure for heterogeneous teams of autonomous vehicles. In: International Conference on Robotics and Automation ICRA 2007. IEEE, Los Alamitos (2007)
14. Deugo, D., Weiss, M., Kendall, E.: Reusable patterns for agent coordination. In: Omicini, A. (ed.) Coordination of Internet Agents, pp. 347–368. Springer, Heidelberg
15. Martins, R., Dias, P.S., Marques, E.R.B., Pinto, J., Sousa, J.B., Pereira, F.L.: Imc: A communication protocol for networked vehicles and sensors, Oceans (2009)
16. Hart, S.G., Staveland, L.E.: Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, Amsterdam (1998)

Enhancing Pilot Training with Advanced Measurement Techniques

Kelly S. Hale[1] and Robert Breaux[2]

[1] Design Interactive, Inc., 1221 E Broadway, Suite 110, Oviedo, FL 32766, USA
[2] Private Consultant, USA
kelly@designinteractive.net, ARA-B@cfl.rr.com

Abstract. Certified Flight Instructors (CFIs) in general aviation are tasked with training student pilots the knowledge and skills related to piloting an aircraft. This requires CFIs to have indepth knowledge about common student errors including early indicators of non-optimal performance in flight, an understanding of probable root cause(s) of non-optimal performance, and instructional techniques to address root cause(s). There is an opportunity to improve CFIs' awareness of common student errors that lead to accidents/incidents and training effectiveness by integrating low fidelity scenario-based training. Such scenarios provided using low cost simulation environments coupled with detailed performance measures outlined in the ADAPT framework can aid CFIs in understanding common errors so that effective recognition and appropriate training intervention is provided to student pilots with the goal of optimizing training while minimizing student accidents/incidents.

1 Introduction

Training focuses on changing cognition, behaviors, and attitudes, where this change is focused on correcting "deficiencies by targeting the right competencies" (Salas et al., 2006). In civilian aviation, certified flight instructors (CFIs) are key personnel in training future pilots, as they are hired to teach students all required knowledge and skills related to piloting an aircraft, and recommending pilots for test flights and certification. While CFIs are provided training handbooks (e.g., FAA-H-8083-9A) and the opportunity to attend training workshops, these resources provide theoretical methods to identify root cause of accidents, instructional theories, and inform CFIs of typical accidents/incidents that occur with student pilots. However, the information available is not organized into a body of instructional techniques based on typical student errors. Information on why accidents/incidents occur must be individually gleaned from National Transportation Safety Board (NTSB) reports or the Aircraft Owners and Pilots Association (AOPA) Nall reports. This information is critical to CFIs, as they must teach recognition/recovery of unsafe flying in order to allow a student to learn from mistakes, but at the same time be ready to take over the controls if an accident/incident is immanent. Oftentimes, the "teaching moment" in flight is not recoverable for safety reasons and precious learning time may lapse between the student error in flight and the opportunity to teach once returning safely to the ground.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 540–545, 2011.

A second challenge faced by CFIs is that student attention is often so focused on flying the airplane that he/she seldom can absorb instruction while in flight. Thus, the typical instructional flying technique is that the instructor demonstrates, then the student emulates; however, "monkey see, monkey do" does not provide convincing evidence of student grasp of the complex relationships of the multitude of factors contributing to safe flight. Diagnostic performance measures are needed to enhance CFI situational awareness of student errors and root cause, as in-flight CFI attention must be shared between attention to student performance while also attending to ensuring the flight environment remains safe. Student observation must come second to maintaining a safe environment.

One obvious solution would seem to be the use of simulation for training. The FAA is a strong proponent of simulator training citing such benefits as “more in-depth training [than] the airplane”, a “very high percentage of transfer of learning”, and “safer flight training” (FAA, 1983). The military and airlines do in fact use simulation extensively, but the general aviation pilot does not have readily available the “free” access to simulators that is afforded to pilots of the airlines. A simulator must fly like the airplane in order for the FAA to approve its use (FAA, 2010), making it expensive to build a simulator (Adams, 2008). One that can be approved for the kinds of flight that represent the typical accident in general aviation has yet to be marketed. Thus, pilots may be turning to unconventional “simulators” from the gaming industry (Parsons, 2010; Beckman, 2009). The question is whether affordable simulation can achieve sufficient benefit to impact positively the accident rate of general aviation pilots. In particular, could the CFI learn to recognize the cues of a student’s eminent accident by using a simulator to re-create typical accident scenarios from Nall Reports (AOPA)? Then, could that learning transfer to the actual aircraft to enhance the student’s training and safety? We believe it is quite feasible.

2 Learning Theories

Two prominent learning theories include (1) Behaviorism, where learning can be observed via measurable responses to stimuli, and (2) Cognitive Theory, which focuses on what happens within the brain – the process of thinking and learning (FAA, 2009). One of the main premises to enhancing performance according to the Behaviorism theory of learning is to provide reinforcement, which is provided by CFIs in aviation training. This theory provides the instructor with ways to manipulate students with stimuli, induce the desired behavior or response, and reinforce the behavior with appropriate rewards. In general, the behaviorist theory emphasizes positive reinforcement rather than no reinforcement or punishment. As an instructor, it is important to keep in mind that behaviorism is still widely used today, because controlling learning experiences helps direct students toward specific learning outcomes.

Cognitive theory focuses more on internal processes related to learning that are not necessarily observable using traditional methods. However, with advances in neurophysiological measurement, cognitive processes are becoming more ‘observable’ utilizing technologies such as eye tracking that can provide quantification of visual attention in real-time and brain-based measurement techniques that can provide indications of cognitive constructs such as workload, distraction, and fatigue. Integrating such measures into scenario-based training opportunities that simulate conditions that occur during flight can provide further insights into a student's performance by not only identifying where performance breakdowns occur, but also provide indications of *why* performance errors occurred. This is particularly relevant for the aviation domain, where multitasking is a key skill required for safe flight. Effectively being able to switch attention and perform multiple tasks simultaneously are two key components student pilots must learn, and thus CFIs must effectively train.

3 Scenario-Based Training for CFI Training

Scenario-based training is one tool that may be implemented to enhance CFI situation awareness and training effectiveness, and this training can be provided via simulation systems that engage students and instructors into the scenario. Parsons (2010, p.36) noted that simulations may be used to "practice and reinforce the lessons learned" from a student perspective. In a similar fashion, scenario-based training could be implemented for CFIs to provide effective learning by developing scenarios that have a clear objective tailored to meet the needs of an instructor, and which capitalize on the nuances of the local environment. Such scenarios could be based on pervious accident reports, which would allow CFIs to view incidents from a number of viewpoints that are not available when they are situated in the cockpit (e.g., can review cockpit view, tower view, from either side of the aircraft), and can also pause and replay incidents. Using such repeated review sessions, CFIs can begin to identify what happened, and identify cues prior to the incident that indicated a potential problem (e.g., too fast in landing approach, excessive control correction).

To further enhance understanding of root cause of incidents, additional advanced measurement techniques beyond those captured by existing desktop-based simulated flight environments may be implemented into a simulated scenario to provide detailed student pilot information. Design Interactive, Inc. has created the Auto-Diagnostic Adaptive Precision Training (ADAPT) framework (Figure 1) that measures, diagnoses, and adapts training based on individualized trainee outcomes. ADAPT is flexible to capture a number of measures simultaneously that indicate trainee state in real-time throughout a training scenario. For example, behavioral data such as flight control manipulations and eye fixations, and brain-based data from electroencephalography (EEG) to indicate cognitive states of workload, engagement, and distraction were incorporated into a flight instrument landing task (Carroll et al., 2010). The suite of measures captured is then analyzed using a diagnostic engine to determine error patterns in behavior and associated root cause(s) related to observed errors. Root cause(s) could include inappropriate scan strategies, non-optimal cognitive states, lack of procedural understanding, etc. Based on the diagnosis, ADAPT can provide precision in one of two ways: (1) adapt the training scenario in real-time to target inefficiencies/deficiencies, and/or (2) provide after action review (AAR) feedback that summarizes trainee performance, key performance inefficiencies/deficiencies, and recommends future training focus.

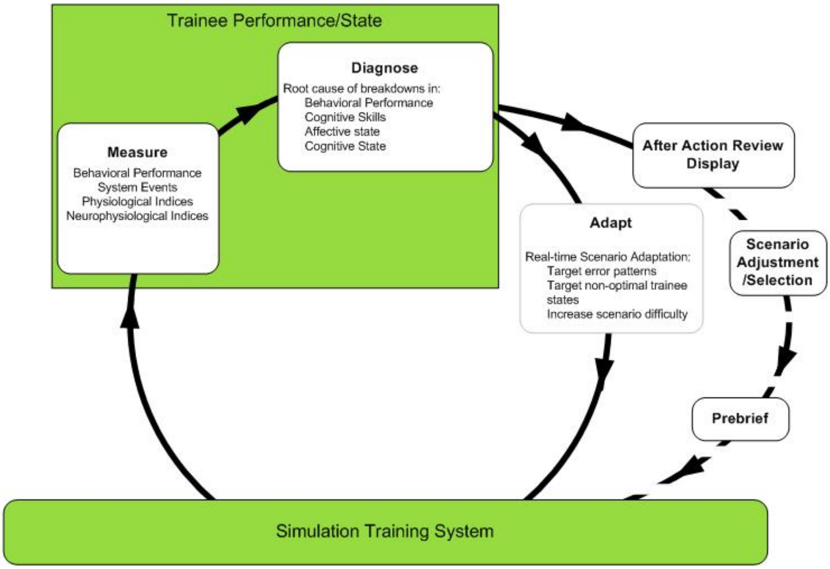

Trainee Performance/State
Diagnose
Root cause of breakdowns in:
Behavioral Performance
Cognitive Skills
Affective state
Cognitive State
Measure
Behavioral Performance
System Events
Physiological Indices
Neurophysiological Indices
After Action Review Display
Adapt
Real-time Scenario Adaptation:
Target error patterns
Target non-optimal trainee states
Increase scenario difficulty
Scenario Adjustment /Selection
Prebrief
Simulation Training System

Fig. 1. ADAPT Framework

For example, ADAPT's AAR components may display eye gaze data showing where a student pilot was visually focused throughout a given time segment of a scenario (Figure 2). This information could provide early indicators of potential accidents that are not otherwise available by observing behavioral outcomes or aircraft maneuvers. For example, it may become evident that students are focused in the cockpit at a specific gauge that is not relevant to the current flight segment and/or where the student inappropriately fixated just over the aircraft nose, which is an early indicator of non-optimal landing techniques that can often lead to hard landings. Insights can be gained not only to errors that happened, but why such errors occurred.

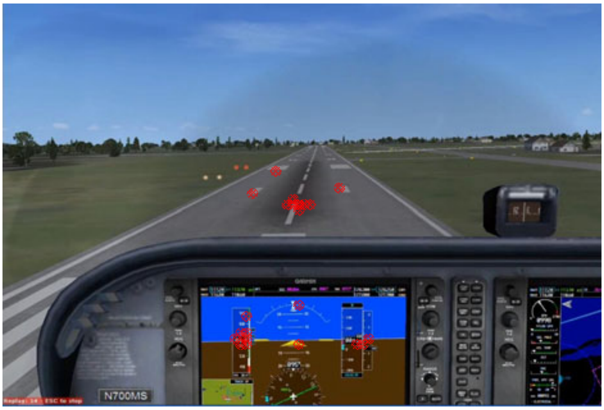


Fig. 2. Fixation Overlay on Approach (red circles indicate fixation points)

The diagnostics based on student behavior could be of great utility to CFIs during their instructor training. By viewing student behavior, including in-depth gaze pattern and cognitive state metrics, CFIs could identify potential causes of accidents, and develop a 'virtual experience database' of underlying causes of common errors (e.g., visual focus in the cockpit or too close to the nose is often cause of porpoising). This knowledge and experience will better prepare CFIs to provide targeted feedback to student pilots that addresses the underlying cause of a poor outcome (e.g., improper scan pattern) as opposed to identifying the poor outcome in isolation (e.g., aircraft porpoise during landing). By focusing on the underlying issue, the student should be better able to adapt their behavior to optimize performance outcomes. Using accident re-creation in a dynamic simulated flight environment, the CFI can observe specific student behaviors (both observable as well as eye tracking data that is 'unobservable' during actual flight) leading to an accident. Such exposure could allow CFIs to temper with detailed knowledge surrounding common student errors, and develop a more thorough understanding of early indicators of non-optimal performance, and identify optimal training intervention techniques and timeline, such as knowing how far to allow students to fly into a mistake for training purposes without risk of a mishap.

Having this detailed knowledge regarding underlying causes that is summarized in a focused, applied manner should provide CFIs a more practical method to learn evaluation skills related to recognizing why errors occur, which allows CFIs to provide targeted feedback to address the underlying cause of inefficient piloting as opposed to simply providing more practice.

4 Future Directions

The CFI is the backbone of General Aviation training. In her introduction, Parsons (2009, p.36) notes that CFIs perform "one of the most vital and influential roles in aviation and, just as in medicine, the work can have life and death consequences. But while the medical profession uses internship and residency programs to provide supervised real world training for newly graduated MDs, newly certificated flight instructors – like new instrument pilots – are mostly left to learn on their own." Systematic exposure to progressively more challenging student behaviors is commonly considered how one gains experience. Parsons (2009) proposes 25 flights in such a progression before the new CFI ever flies with a beginning student. Simulation could be used to expose the CFI in a safe environment to student behaviors that have resulted in accidents, and thus better prepare the CFI for successful accomplishment of their vital role.

Further, the ADAPT framework could be extended from a student-focused evaluation to a CFI-focused evaluation. Specifically, performance measures collected and summarized from a simulated flight segment may focus on CFI time to identify student error, CFI method of instruction for addressing error, time for student pilot to recover from error, and CFI workload during flight segment. Providing such detailed feedback regarding their performance in instructing a pilot while maintaining flight safety provides unprecedented training opportunities to CFIs that are currently unavailable. By implementing such training techniques, CFI situation awareness of student pilots and their own instructional abilities can be explicitly measured and targeted feedback can be provided to improve instructor-pilot interaction, student training, and accident/incident rates in civil aviation.

References

1. Adams, R.: Aircraft Source Data: High Price Still a Concern. The Journal for Civil Aviation Training, (3) (2008), http://cat.texterity.com/cat/2008-3/#pg16 (Viewed January 20, 2011)
2. Beckman, W.S.: Pilot Perspective on the Microsoft Flight Simulator for Instrument Training and Proficiency. International Journal of Applied Aviation Studies 9(2), 171–180 (2009)
3. Carroll, M., Fuchs, S., Hale, K., Dargue, B., Buck, B.: Advanced training evaluation systems: leveraging neuro-physiological measurement to individualize training. In: Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Conference Proceedings (2010)
4. FAA. Airplane Simulator and Visual System Evaluation. Advisory Circular AC 120-40 (1983)
5. FAA. Aviation Instructor's Handbook 2008: FAA-H-8083-9A. Aviation Supplies & Academics, Incorporated: Newcastle, WA (2009)
6. FAA. Flight Simulation Training Device Qualification Guidance (2010), http://www.faa.gov/about/initiatives/nsp/flight_training/ac/ (updated: 11:55 am ET August 11, 2010), http://www.faa.gov/about/initiatives/nsp/flight_training/ac/media/120_40.pdf
7. Parsons, Susan (eds.): Best Practice for Mentoring in Aviation Education (2009), http://www.faa.gov/training_testing/training/media/mentoring_best_practices.pdf
8. Parsons, Susan (eds.): Runway. FAA Safety Briefing. (September/October 2010), http://www.faa.gov/news/safety_briefing/2010/media/SepOct2010.pdf
9. Salas, E., Priest, H.A., Wilson, K.A., Burke, C.S., Adler, A.B., Castro, C.A., Britt, T.W.: Scenario-based training: improving military mission performance and adapbility. In: Britt, T.W., Adler, A.B., Castro, C.A. (eds.) Military Life: The Psychology of Serving in Peace and Combat, Praeger, Westport (2006)

Rule Fragmentation in the Airworthiness Regulations: A Human Factors Perspective

Don Harris

HFI Solutions Ltd., Bradgate Road,
Bedford, MK40 3DE, United Kingdom
don.harris@hfisolutions.co.uk

Abstract. Human error has been identified as the primary risk to flight safety. Two of the more pervasive aspects of Human Factors encountered throughout the airworthiness regulations are error and workload. However, as a result of increasing organizational inter-dependence and integration of aircraft systems it is argued that the manner in which these issues are addressed in the aviation regulations is becoming increasingly incompatible with human and organizational behavior in an airline. Workload and error are both products of complex interactions between equipment design, procedures, training and the environment. These issues cannot be regulated on a localized basis. A more systemic, holistic approach to Human Factors regulation is required. It is suggested that a Safety Case-based approach may be better used as an adjunct to existing regulations for Human Factors issues.

Keywords: Regulations, Workload, Error, Accidents, Socio-technical systems, Safety Case.

1 Introduction

For the last decade the serious aircraft accident rate has remained relatively constant at approximately one per million departures [1]. However, as engineering integrity has improved, the proportion of accidents resulting human error has increased. In over 75% of cases the actions of the crew have been identified as a major contributory factor [2] making human error the primary risk to flight safety.

Human Factors in aviation is intimately associated with the pursuit of safety. It is embedded in selection, training and design processes and is a cornerstone of all safety management systems. Furthermore, 'good' Human Factors practice is mandated (either implicitly or explicitly) via many airworthiness regulations. This paper examines the treatment of just two of the more pervasive aspects of Human Factors encountered throughout the airworthiness regulations: error and workload.

The roots of human error are manifold and have complex inter-relationships with all aspects of the operation of a modern airliner. For example, during the last decade, as a result of a series of accidents involving highly automated aircraft (e.g. the accident involving an Airbus A320 at Strasbourg and the Boeing 757 near Cali) 'design induced' error was of particular concern to the airworthiness authorities. The

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 546–555, 2011.

Federal Aviation Administration (FAA) study of the pilot-aircraft interfaces in highly automated aircraft [3] contained criticisms relating to aspects such as autoflight mode awareness/indication; energy awareness; confusing and unclear display symbology, and a lack of consistency in Flight Management Systems. In 1999 the Department of Transportation tasked the Aviation Rulemaking Advisory Committee to '*review the existing material in FAR/JAR 25 and make recommendations about what regulatory standards and/or advisory material should be updated or developed to consistently address design-related flight crew performance vulnerabilities and prevention (detection, tolerance and recovery) of flight crew error*'. In Europe this has resulted in a new airworthiness rule (CS 25.1302). The US FAA will soon follow in adopting this rule. However, many errors have their root causes in a range of other aspects of the operation of the aircraft, not just flight deck design.

Mental workload assessment has been a component of the flight deck certification process since 1993. Indeed, until the recent implementation of the flight deck certification requirement aimed at avoiding design induced error [4] the assessment of pilot workload was the primary rule associated with Human Factors. Appendix D to FAR/CS 25.1523 and FAA Advisory Circular AC 25-1523-1 define six basic workload *functions* and ten workload *factors* [5]. Workload *functions* are related to the basic tasks of flying the aircraft (e.g. flight path control; navigation; communications): these facets impose workload on the pilots. The workload imposed by these *functions* can be either ameliorated or exacerbated by the workload *factors*. These are aspects that relate to the design of the aircraft and/or its operation, such as the accessibility, ease, and simplicity of operation of all necessary flight, power and equipment controls; the extent of required monitoring of systems, and the degree of automation provided in the aircraft systems to afford automatic crossover to, or isolation of difficulties after failures or malfunctions.

Workload is a stressor that needs controlling. The workload *factors* can all be managed by various aspect of good design. In this context, aircraft certification is concerned with the measurement of the workload imposed by the aircraft and its operation to demonstrate that it is within acceptable bounds for safe fight.

2 Integration and Interdependency

Applegate and Graeber [6] described the increasing levels of integration and interdependency of aircraft systems in Boeing jet transport aircraft. Early airliners such as the Boeing 707 and 727 had relatively independent systems managed by a Flight Engineer. The initial Boeing 737s had simplified systems with greater levels of automation for the management of systems to allow two crew operations. Nevertheless, the aircraft still utilized analog technology. Later Boeing 757/767 models were the first to use digital technology. However, their basic architectures were simply digitized versions of earlier analog systems with little integration. The Boeing 777 employed new system architectures with greater use of digital technology. The Boeing 777 possesses highly integrated systems with inputs providing data for a variety of aircraft functions. This increased complexity and integration, though, also impacted upon the system design.

In response to these higher levels of system integration, several industry/ government teams developed corresponding safety requirements and practices (e.g. SAE ARP4754 ‘Certification Considerations for Highly Integrated or Complex Airplane Systems’ [7] and SAE ARP4761 ‘Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems’ [8]). However, the airworthiness regulations themselves, as contained in Code of Federal Regulations, Title 14 (Aeronautics and Space); Part 25 (Airworthiness standards: Transport Category Airplanes) still continue to adopt a ‘system-by-system approach’. Aircraft systems (or perhaps more properly now, ‘functions’) are still considered largely on an engineering, standalone basis with little consideration for their integration.

In addition to aircraft becoming more integrated, airline operations have also become more integrated. Consider the turn-round operation which has traditionally been viewed as a standalone process with responsibilities shared between the airline and airport. Emphasis in operations is now becoming placed upon synchronizing all stakeholders. ATM (Air Traffic Management) links the arrival, turn-round and departure phases as one entity. The associated ground processes and en-route traffic are now considered as part of a time-dependent chain. Airport Collaborative Decision Making (CDM) is used as a mechanism to integrate airports into the ATM network. The CDM turn-round process includes airport operator, airline, air traffic control, ground handling and Central Flow Management Unit. Flight update messages and departure planning information are in place to inform all participating CDM partners about a particular flight’s progress [9]. In addition, the nature of the airline business has changed dramatically. There is now a great deal more outsourcing and sub-contracting of functions previously undertaken within an airline. Organizationally, airlines are now semi-‘open’ systems (in terms of Systems Theory [10]). To illustrate, airlines operate into a wide range of airports; maintenance is often provided by third parties and ATM/ATC is provided by the national authorities of the countries which they either operate into or overfly. Furthermore, some low-cost airlines may not even own their aircraft, employ their own ground and check-in personnel, and in extreme cases, may not even employ their own pilots [11].

However, as a result of airlines simultaneously becoming more organizationally ‘open’ while also exhibiting a much higher degree of integration of operations, it has become easier for errors to promulgate between organizations [12]. For example, the error proximal to the accident in the Uberlingen mid-air collision was a failure of the Skyguide air traffic controller in Zurich Air Traffic Control Centre to notice that two aircraft were on converging flight paths. This error was then compounded and promulgated across organizations when the controller also gave incorrect positional information concerning the conflicting Boeing 757 to the crew of the Tupolev Tu-154M when they expedited their descent: he also failed to notice that the Boeing had initiated a descent in response to a TCAS advisory. These errors were partially a result of his workload being high because he was the only controller on duty and he was overloaded because he was simultaneously trying to coordinate the approach of an Airbus A320 into the nearby Friedrichshafen airport.

3 Regulatory Framework

In many respects, the airworthiness regulations addressing Human Factors issues are extremely fragmented. For the sake of this discussion the US regulatory structures will be used to illustrate, as covered in Part 25 of the Code of Federal Regulations (CFR), Title 14 (Aeronautics and Space) [13].

Over the last 50 years, so as to accommodate all the different facets of airline operations, the commercial aviation system has developed a rule system has become increasingly diverse and complex. For example, the basics of pilot licensing and training are covered Part 61 of the Federal Aviation Regulations (FARs). These are supplemented by further license endorsements to fly an aircraft at night, fly an aircraft with more than one engine, fly in controlled airspace, etc. To fly fare-paying passengers requires an Airline Transport Pilots certificate. The regulations covering training technology (e.g. flight simulators) are covered in Part 60 of the regulations. The basic rules for the operation of aircraft are covered in FAR Part 91. FAR Part 119 applies to the operation of a civil aircraft as an air carrier or commercial operator. This specifies the management roles and processes required for an Air Operator's Certificate (largely organizational issues). The manner in which an airline's operations are conducted is specified in Part 121, 125 and/or 135 (further organizational and operational matters).

From a Human Factors perspective, Part 25 deals with the flight deck interfaces, which are covered in a number of separate regulations; as noted earlier licensing, training and the technology of training are covered Parts 61 and 60. These training-associated parts of the regulations are all generic requirements but once an aircraft weighs over 12,500 lbs a specific type rating is required which ensures that there is a good 'fit' between the aircraft, and the pilot's skills, knowledge and ability to fly it (a product of training). Organizational structures and function are dealt with in Parts 119, 121, 125 and 135. This can be further illustrated by superimposing the various parts of the regulations over a simple representation of a classical 'Perception-Decision-Action-Feedback' loop which describes a simple manual flying task (see Figure 1). The aircraft controls and displays are part of the aircraft and hence are regulated in Part 25. The basic skills required to fly an aircraft and their assessment is covered in Part 61: the flight simulator technology to inculcate these skills is considered in Part 60. However, how the pilot uses these components in an airline context (i.e. how the task of flying an aircraft containing passengers and cargo is undertaken) and the 'fit' between the aircraft and pilot is regulated in Parts 121/125/135. Flying the airplane in a safe and appropriate manner within the air traffic system is covered in Parts 91 and 119 and the wider environmental context of operations (not considered in Figure 1) includes yet more parts of the regulations, such as Part 71 (Designation of Class A, B, C, D, and E Airspace Areas; Air Traffic Service Routes; and Reporting Points); Part 77 (Objects Affecting Navigable Airspace); Part 139 (Certification of Airports) and Part 153 (Airport Operations). This brief description merely begins to scratch the surface of the complexity and fragmentation of the regulations. However, the one thing that the rules are not explicitly concerned with is the *system* of transporting people and cargo safely from A to B, despite this being their intent.

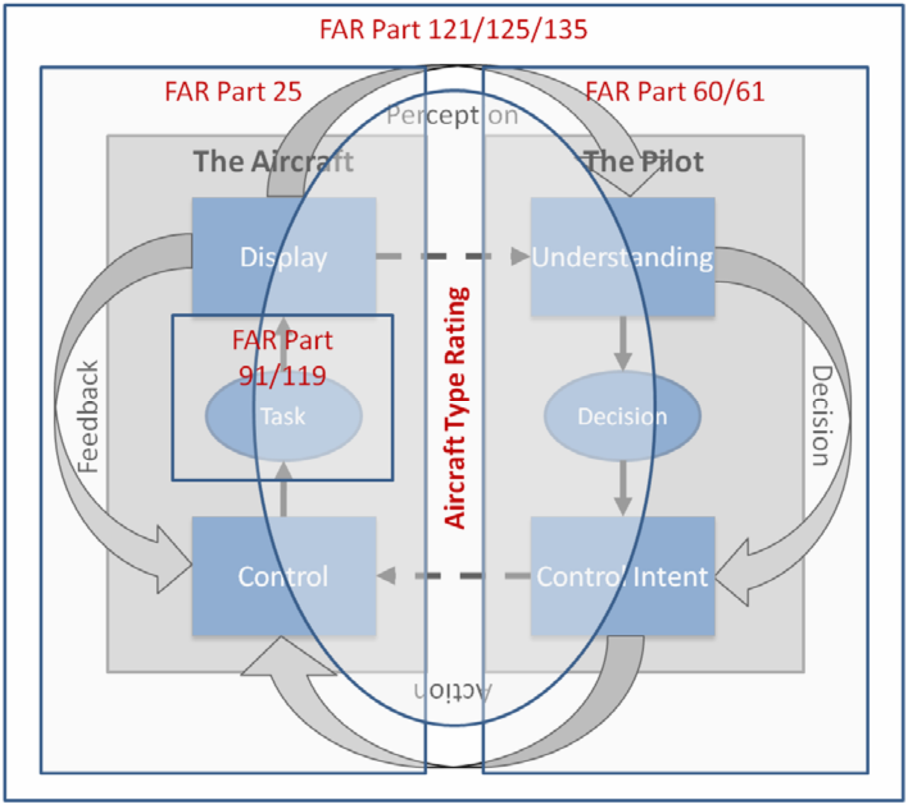


Fig. 1. The concept of the Human-Machine (Flight Deck) Interface superimposed over a representation of the classical 'Perception-Decision-Action-Feedback' control loop with the various parts of FARs further superimposed over the diagram to illustrate the fragmentation of the regulatory system (adapted from Harris [14])

4 Systemic Nature of Workload and Error

Pilot workload is a product of the number and difficulty of the tasks to be performed in the time available; the usability of the flight deck equipment and the interactions with the flight task and other stressors. The topic of pilot workload appears in 37 different FAA Advisory Circulars relating to 19 separate parts of the regulations. The system objective should be to manage pilot's workload but this is handled in different ways across the various parts of the rules. Most often, the regulatory requirement is simply that the component/function under consideration should not impose unduly high levels of workload; but how can this be achieved without the wider consideration of other aspects, such as the procedures involved; design of the flightdeck equipment; the other tasks being performed simultaneously and the environmental context?

Error in the operation of large commercial aircraft appears in 45 FAA Advisory Circulars across 24 parts of the regulations all addressing different parts of the socio-technical system of operating an airliner. However, the systemic view of error is that it is a product of equipment design, procedures, training and the environment [15]. It has also been described how errors can promulgate across organizational boundaries. Error has its roots in the surrounding socio-technical system. Again, the system objective is to manage error but there is no regulatory systemic approach to the eradication, control and management of error (the classical error 'troika').

When the consideration of these issues is described in this manner it becomes apparent that the regulatory structures impede making system-wide improvements in safety and efficiency. If workload and error have a system-wide etiology, they must be regulated collectively across the many separate aspects of the system of regulation if they are to be tackled efficiently and effectively. It may be the case that it is the regulatory system itself that is preventing further improvements in safety (hence the observation that during the last decade the serious accident rate has plateaued at approximately one per million departures). Simply adding another local regulation to fix one specific aspect of a much wider system problem is unlikely to have any major effect. Fragmented rules that do not adopt a system-wide perspective may not increase safety to the degree anticipated.

Many Human Factors issues lie not within an individual regulation but between regulations. The new European Human Factors flight deck certification rule (CS 25.1302[5]) tries to take a task-based approach but is limited by the scope of Part 25 itself (which addresses only the design of the aircraft). Factors outside Part 25 cannot be considered when assessing compliance, nor is it permitted that the regulation can address issues outside those associated with the design and structure of the aircraft. While the probability of design-induced error on the flight deck may be significantly reduced after implementation of this rule, the level of overall risk in flight operations and the accident rate may only be marginally decreased as a result of this failure to adopt a systemic perspective: not all errors on the flight deck fall into the category of 'design induced'. In fact, to suggest that there is merely a single source to any error is over-simplistic. To re-iterate, error and workload are products of interactions between the pilots, aircraft, procedures and the environment. The notion of human error having a single root cause is an oversimplified view of the roots of failure. Furthermore, flying an aircraft progresses on a task-by-task basis *not* a system-by-system basis (the approach implicit in the structure of the regulations). The regulatory structures are not commensurate with human performance.

Consider the following case study of the Singapore Airlines Flight 006 accident, in Taipei, Taiwan. Workload and error were both involved in the sequence of events but the source and/or control of these factors cannot be isolated within any one single part of the regulations. Furthermore, error can result from factors external to the aircraft (but which are still covered within the wider regulatory structures).

4.1 Singapore Airlines Flight 006, Boeing 747, Taipei, Taiwan, 2000

Flight SQ006 crashed on departure from Taipei airport at night in heavy rain and strong winds from a passing typhoon. The accident was attributed to a lack of situational awareness which resulted in the crew erroneously taking off from the wrong runway.

The crew was in a hurry to depart before the weather deteriorated further, closing the airport. They were cleared for departure on runway 05L as runway 05R was closed between taxiways N4 and N5 owing to construction work; as a result, runway 05R was re-designated as taxiway NC. The wind was reported as 36 knots (gusting to 56 knots) with a runway visual range of 600 meters. Upon reaching the end of the taxiway the crew turned right into taxiway N1 and immediately made a 180-degree turn onto runway 05R. After very short hold SQ006 started its takeoff roll. Just after V1 (the go/no-go decision speed) the aircraft hit concrete barriers, excavators and other construction equipment, crashing back onto the runway and then breaking up and bursting into flames [16]. Seventy-nine passengers (out of 159 on board) and four of the 20 crew died.

The reasons contributing to the decision to takeoff from the wrong runway were attributed to a variety of causes, including: poor CRM (the crew did not ensure they understood the correct route to runway 05L and no one confirmed which runway they had entered); there was misleading runway/taxiway lighting leading onto runway 05R resulting in the Captain focusing his attention on following these taxiway centre-line lights; crew workload was much higher than normal as a consequence of the inbound typhoon; and the environmental conditions were poor (strong crosswind, low visibility and slippery runway). There was information available on the flightdeck suggesting that the aircraft was on the incorrect runway (e.g. the runway edge lighting not illuminated; there were lighting configuration and width differences between Runway 05L and 05R; and the para-visual display indicated that the aircraft was not aligned with the runway localizer) but these factors were ignored.

In this example aspects of CRM; pilot training; airport design and operation; environment (weather factors imposing workload); Air Traffic Control and flightdeck display design were all implicated in the sequence of events leading to the accident. If examined using the FAA regulatory structures, Parts 60 (Flight Simulation Training Device Initial and Continuing Qualification and Use) and 61 (Certification: Pilots, Flight Instructors, and Ground Instructors) were implicated in the training of the pilots. The oversight of aircraft operations would be subject to overview under Part 135 (Operating requirements: Commuter and on demand operations and rules governing persons on board such aircraft). The regulation of the airport itself would be covered under Parts 139 (Certification of Airports) and 153 (Airport Operations). The design of the aircraft flightdeck and its equipment is covered under Part 25 (Airworthiness standards: Transport Category Airplanes). However, the errors made and the workload imposed on the crew cannot be attributed to any one single factor (hence any one part of the regulations). To re-iterate, the causes of workload and error are systemic.

5 A Safety-Case Approach

The current aviation safety-related regulatory structure has evolved over five decades, or more. It began when engineering considerations took precedence and when aircraft systems were relatively independent. As the reliability and structural integrity of aircraft has improved, human error has become the primary risk to flight safety. However, in recent years the serious accident rate has remained relatively constant at approximately one per million departures. It has been suggested that this plateau in the accident rate may be at least partially attributable to the fragmentary nature of the regulations when dealing with Human Factors. The structure of the aviation-safety regulations is not compatible with human behavior which progresses on a task-by-task basis not on a system-by-system basis.

Furthermore, aviation accidents rarely have a single error or cause underlying them:

> *'…it is well established that accidents cannot be attributed to a single cause, or in most instances, even a single individual. In fact, even the identification of a "primary" cause is fraught with problems. Instead, aviation accidents are the result of a number of causes...'* (Shappell and Wiegmann, p. 60 [17]).

The regulations were also developed at a time when airline operations were much simpler, lower tempo and were less integrated. During this period airlines were also more organizationally -'closed'. However, safe working practices are dependent upon the control management exercises over work processes and factors external to the organization.

Fragmented rules that do not afford a system-wide perspective may not increase safety to the degree anticipated. From a Human Factors perspective a coherent link between aircraft design, training and operations is required to enhance both safety and efficiency that is also commensurate with human behavior. Rules and regulations need to be future proof, defined in terms of the required result not the method to achieve it. Airworthiness rules that are too prescriptive may stifle technological and operational innovation and also potential advances in safety.

Generation of an operational Safety Case may be a regulatory approach which satisfies the requirements of these criteria and is compatible with the nature of both human behavior and modern airline operations. Safety Cases are commonly used in the UK offshore oil and gas industry. Each installation must demonstrate (to the UK Health and Safety Executive) how major accident hazards are adequately controlled and that the management system is suitable. This approach to Safety Management was mandated after the accident to the Piper Alpha oil production platform in the North Sea in 1988 that killed 167 personnel [18]. At this time the offshore multinational companies operated the installations largely with their own personnel. However, during the last twenty years oil companies have restructured and in common with the airline industry, sub-contracting has become commonplace.

The Safety Case is a structured argument, supported by a body of evidence that provides a comprehensive and valid case that a system is safe for a particular type of operation in a particular operating environment. From a Human Factors perspective specific topics under consideration in safety case presentations normally cover the competencies required to perform the work; training and training needs analysis; development and maintenance of procedures; communication processes; manning levels; automation and allocation of function; supervision of staff; shift patterns; hardware and software layout; environmental performance shaping factors; human error potential and safety culture. Safety Cases are not prescriptive: the aim is to demonstrate systems meet the required safety goal; they do not separate the human from the system; they are evidence-based and are subject to continual revision (they change in response to changes in the nature of operations). This approach is also becoming used much more frequently in defense aerospace, for example in the Eurofighter Aircraft Avionics project; the BAe Hawk Aircraft Safety Justification and in Military Air Traffic Management Systems. A similar approach is being use for the safety evaluation of civil Unmanned Air Systems [19]. Furthermore, the basis for safety cases is already being used by all airlines as part of their Safety Management Processes.

If safety regulation is to progress in a manner compatible with the management of workload and error it has to progress on a systemic basis, not a system-by-system basis. A Safety Case-based approach provides this opportunity. This is not to say that it should replace the current set of regulations as this would be completely impractical. However, it can be used as an adjunct and/or alternative where a suitable waiver to existing regulations is granted.

References

1. Boeing Commercial Airplane Group: Statistical Summary of Commercial Jet Airplane Accidents Worldwide Operations 1959–2008. Boeing Commercial Airplane, Seattle, WA (2009)
2. Civil Aviation Authority: Global Fatal Accident Review 1997–2006 (CAP 776). Civil Aviation Authority, London (2008)
3. Federal Aviation Administration. Report on the Interfaces between Flightcrews and Modern Flight Deck Systems. US Department of Transportation, Washington, DC (1996)
4. Federal Aviation Administration: Advisory Circular AC 25-1523-1 Minimum Flight Crew. US Department of Transportation, Washington, DC (1993)
5. European Aviation Safety Agency: Certification Specifications for Large Aeroplanes (CS- 25): Amendment 7. EASA, Cologne, `http://www.easa.europa.eu/ws_prod/g/rg_certspecs.php#CS-25`
6. Applegate, J.D., Graeber, R.C.: Integrated safety systems design and human factors considerations for jet transport aeroplanes. Human Factors and Aerospace Safety 1, 201–221 (2001)
7. Society of Automotive Engineers: Certification Considerations for Highly Integrated or Complex Airplane Systems (SAE ARP4754). Society of Automotive Engineers, Warrendale, PA (1996)
8. Society of Automotive Engineers: Guidelines and Methods for Conducting the Safety Assessment Process on Civil Airborne Systems (SAE ARP4761). Society of Automotive Engineers, Warrendale, PA (1996)
9. Groppe, M., Pagliari, R., Harris, D.: Monitoring the Aircraft Turn-Round Process: Applying a Qualitative Cognitive Model Based on Field Observations. 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, September 20-24, pp. 266-277. European Association for Aviation Psychology, Amsterdam, NL (2010)
10. von Berthalanfry, L.: General systems theory: general systems. Yearbook of the Society of General Systems Theory, 1, 1-10 (1956)
11. Harris, D.: Keynote Address: An Open Systems Approach to Safety Culture: Actions, Influences and Concerns. In: Australian Aviation Psychology Association (AAvPA) International Conference – Evolving System Safety, Sydney, Australia, November 9–11. Australian Aviation Psychology Association, Victoria (2006)

12. Harris, D., Li, W.-C.: An Extension of the Human Factors Analysis and Classification System (HFACS) for use in Open Systems. Theoretical Issues in Ergonomics Science (in press)
13. Code of Federal Regulations Title 14: Aeronautics and Space. Part 25: Airworthiness standards: Transport Category Airplanes. National Archives and Records Administration, Washington DC, http://www.gpo.gov/nara/cfr
14. Harris, D.: Human Factors for Flight Deck Certification: Issues in Compliance with the new European Aviation Safety Agency Certification Specification 25.1302. Journal of Aeronautics, Astronautics and Aviation. Series A 42, 11–20 (2010)
15. Dekker, S.W.A.: The Re-Invention of Human Error. Human Factors and Aerospace Safety 1, 247–266 (2001)
16. Aviation Safety Council: Aircraft Accident Report: Crashed on a Partially Closed Runway During Takeoff. Singapore Airlines Flight 006 (Boeing 747-400, 9V-SPK); CKS Airport, Taoyuan, Taiwan, October 31, 2000. ASC, Taiwan, Republic of China (2002)
17. Shappell, S.A., Wiegmann, D.A.: Applying Reason: the Human Factors Analysis and Classification System (HFACS). Human Factors and Aerospace Safety 1, 59–86 (2001)
18. Cullen, The Honourable Lord: The Public Inquiry into the Piper Alpha Disaster. HM Stationery Office, London (1990)
19. Civil Aviation Authority: Unmanned Aircraft System Operations in UK Airspace – Guidance (CAP 722). Civil Aviation Authority, London (2010)

Development of a Reconfigurable Protective System for Multi-rotor UAS

Thomas Irps, Stephen Prior, and Darren Lewis

Department of Product Design and Engineering,
School of Engineering and Information Sciences, Middlesex University,
London N14 4YZ, United Kingdom
tirps@yahoo.com

Abstract. The purpose of this study is to illustrate how the design and deployment of a minimal protective system for multi-rotorcraft can cater for changes in legislation and provide for greater use both in and outdoors. A methodology is presented to evaluate the design and development of a system which protects both single axial and co-axial rotorcraft. The key emphasis of the development presented is the scenario in which the multi-rotorcraft can fly with increased speed including the capability of flying through windows and doors without the fear of system failure due to rotor disruption. Discussed as well is the degree of autonomy the reconfigurable system should feature as well as the effects of drag and added component mass to the performance of the system.

Keywords: Autonomous system, landing gear, reconfigurable system, unmanned aerial vehicles.

1 Introduction

In recent years the amount of research and development in Unmanned Aerial Systems (UAS) has grown substantially due to the shift in investment and policy of major military industrialized nations. There are some 100 U.S. companies, academic institutions, and government organizations developing over 300 UAS designs in the U.S. alone [1]. In 2008 the international trade association for unmanned aircraft had 1,400 members in 50 member states [2].

Reviewing the latest US UAS Roadmap 2010 - 2035 indicates that the military aerial strike force will equate to 50% manned and 50 % unmanned aircraft [3]. This creates a large investment opportunity and the amount of new Small to Medium Enterprises (SME) in the UAS market are continuously growing. Services are not only limited to the military or law enforcement agencies. The technology is also filtering down to both the hobby enthusiast and new commercial enterprises such as that of photography and video production. Forums such as diydrones.com illustrate the possibility, quantity and the level of maturity that the hobbyist UAS market is achieving.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 556–564, 2011.

Specific to the development featured in this paper we will consider the multi-rotor range of Vertical Take-Off and Landing (VTOL) UAV´s. These systems can achieve four degrees of freedom (X, Y, Z, and RZ), it also features the ability to hover and perch.

1.1 FAA and CAA

With the growth of Unmanned Aerial Systems new problems have arisen such as mid-air collision, air space regulation, user registration, national security and health and safety. In order to deal with many of these facets the US Federal Aviation Administration (FAA), UK Civil Aviation Authority (CAA) and more specifically the European Aviation Safety Agency (EASA) created new regulations to deal with these problems. The regulations now include systems under the 7 kg bracket as illustrated in Table 1. Vertical Take Off and Landing systems are incorporated into these regulations and as such all vehicles have to be registered. By definition all aerial vehicles including those found in toy shops should fall under the requirements of user and aircraft registration with the CAA in the UK.

Table 1. CAA Weight Classification table [4]

Weight Classification Group	Civil Category	Mass (kg)	Broad Military Equivalent	Civil Regulation
1	Small Unmanned Aircraft	20 or less	Micro (<5Kg) Mini (<30Kg)	National
2	Light UAV	More than 20 to 150	Tactical	
3	UAV	More than 150	MALE/HALE	EASA

As an example, the Unites States National Air Space encompasses an average of over 100,000 aviation operations per day, including commercial air traffic, cargo operations, business jets, etc. [5]. Through the addition of UAS the number of registrations and the quantity of airborne vehicles will greatly increase.

1.2 Problem Statement

The CBP accident rate is 52.7 accidents per 100,000 flight hours (the standard safety data normalization factor/the standard on which safety data is reported). This accident rate is more than seven times the general aviation accident rate (7.11 accidents/100,000 flight hours) and 353 times the commercial aviation accident rate (0.149 accidents/100,000 flight hours) [5].

Studies focused on the cause of UAS accidents illustrate the need for regulations as well as safety system development due to the high rate of human errors [6] [7].

One of the major potential hazards of multi-rotorcraft are its exposed blades. With the brushless motors rotating in excess of 9000 rpm and the propellers featuring sharp edges this can produce deep cuts to exposed skin. With the addition of the system flying at speeds above 3 m/s can and weighing up to approximately 5kg including the rotating blades can produce serious health and safety issues.

The second issue is with the impact survivability of the system. If a multi-rotor UAV were to be deployed and the rotor would be the target of inbound objects or collides with the surrounding architecture the potential for a system failure is high. In order to provide a more durable and reliable system a physical protection method is required.

1.3 Deployment Scenario

Currently multi-rotorcraft are not deployed in military tactical missions abroad due to their short flight endurance and system survivability but have been found useful in law enforcement scenarios such as crowd control at demonstrations. Military research remains active due to the potential capability of the multi-rotor UAS. In law enforcement the Merseyside police force apprehended the first criminal using a Quad rotor featuring first person view (FPS) capability [8]. Kent police and BAE systems have been trialing such systems for deployment at the 2012 Olympic games in London, UK [9] [10].

1.4 Existing Protection Methods

The current method of protection which is provided in commercial systems is that of a fixed enclosure which protects the surrounding environment from the rotating blade. An example of this is found in the AR.Parrot Drone [11]. But with such a protection system only a very specific model and design of multi-rotor can be used. These commercial systems only cater for very small payloads which ultimately leads to a single choice between sensor payload or protection system.

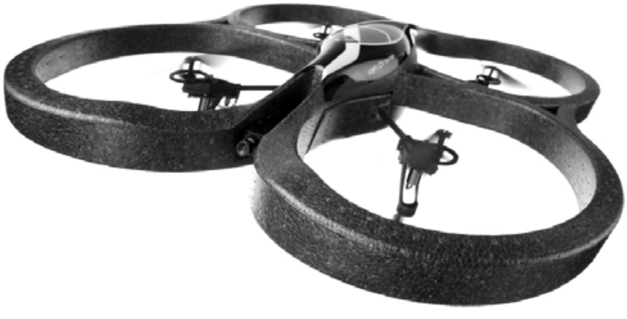


Fig. 1. AR Parrot drone protective enclosure [11]

1.5 Identifying the Key Development Aspects

The key audience and consumer are hobbyists, professionals and developers. The key areas to develop are identified as:

- A system that caters for a variety of systems ranging from Tri-Rotors to Octo-Rotors
- A system that caters for a variety of different propeller and motor dimensions

- A system that does not reduce the field of view of attached cameras
- A lightweight system with low addition of drag
- Which allows the system to withstand impact at different angles and speeds

2 Development Criteria

To validate the development a weighted matrix was used to evaluate all designs and mechanisms which led to three final development routes with one final prototype. The criteria used was that of the following:

Table 2. Categories for the evaluation and development of the landing gear

Categories	Weighting
Horizontal Impact Survivability	1.5
Mass	1.3
Vertical Impact Survivability	1.2
Modularity	1.2
Landing Stability	0.9
Multifunctional	0.9
Field of View	0.9
Simplicity	0.9
Portability	0.7
Costing	0.5
Total	10
Maximum possible score	100

Horizontal Impact Survivability. The system has to be able to survive horizontal impact at defined velocities. These velocities are initially set at 2 m/s for both axis and will be increased accordingly. A flat surface is considered for the horizontal impact.

Mass. The current limit of the landing system as whole should not exceed 150 grams.

Vertical Impact Survivability. The system has to be able to survive vertical impact at defined velocities. These velocities are initially set at 2 m/s for both axis and will be increased accordingly. A flat surface is considered for the vertical impact. The landing system should have a large contact area or contain a flexible structure to absorb the impact.

Modularity. The objective is to create a system that is not bound to one design and that the system only requires as many landing gears as it features motor and propeller combinations.

Landing Stability. The system has to have the ability to cope with landings at an angle to the surface. This maximum set capable angle is of α (pitch) 0-30°, β (roll) 0-30° and the combination of both. We assume that the landing surface is flat.

Multifunctional. The system should provide the ability to reconfigure from one configuration to another rather than be providing both at the same time, thus reducing mass and drag.

Field of View. The system should provide for an un-obstructive field of view for the sensor payload whilst airborne.

Simplicity. This criteria is a collection of various factors which include ease of manufacture, serviceability, ease of assembly, reduction of the number of components, passive rather than active solution.

Portability. The system should be capable of disassembly and be "backpackable" which means that it should fit within the standard issued backpack for infantry soldiers. The main dimensions to be considered is volume which the maximum is set at 2 litres.

Costing. The more cost effective the solution the better. The current target is set at £100 with a maximum rate set at £140.

2.1 Single Axial System

The single axial rotorcraft features a system that can be attached beneath each individual motor and is controlled via a central control board. The current mass for each individual system is of approximately 60 grams including mechanism drive motor. With the brushless motors running at 14 V and at full thrust the simulated decrease in performance is estimated to be of 5.4% due to the addition of both mass and drag.

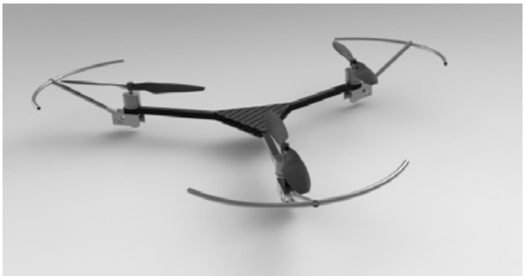


Fig. 2. Preliminary outline of how the single axial solution is placed

2.2 Co-Axial System HALO™

For the Co-axial solution the development was based around the ASL HALO™ Co-Axial Tri-Rotor [12]. Rather than re evaluating the supporting structure of the motor attachments the system was designed in such a manner that it could be installed by replacing structural elements. The key is placing the mechanism in between both motors which requires a different method of a self locking mechanism. For the first iteration, all three landing gears together weigh approximately 327 grams. Running the motors at 14 V and full thrust capability would result in an estimated decrease of simulated performance of approximately around 4.9% due to drag increase.

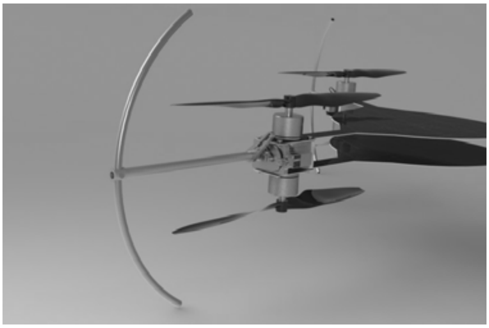


Fig. 3. Preliminary outline of how the co-axial solution is placed

2.3 Mechanism Design

To reduce the overall count of components involved and make the system inherently strong a self-locking mechanism is required for both positions. This mechanism has to also require low amounts of torque to reduce the operating power. The advantage of the mechanism is to provide a effective landing platform as well as a rotor protection. Thus reducing mass, increasing the capability of stronger impact resistance and reducing the volume of components.

2.4 Control Infrastructure

The system is controlled via a microprocessor which is operated either independently through height measurement, linked to the remote operator or in conjunction with the flight operating system. The microprocessor will primarily operate the individual motors but also verify if each motor is in the locked position according to the prescribed condition of flight or take-off and landing. One autonomous method to do this would be to feature a height measurement sensor and consequently defines the status of the system. The second method is that of remote user control which dictates what position is required. The final solution is used in conjunction with the flight processing unit which can evaluate uncontrolled flight and decent, i.e. in case of an interruption in the signal. What still needs to be reviewed is what part of the system requires more protection in case of uncontrolled descent. Is it the expensive sensor payload, such as thermal imaging cameras or is it the UAS itself. The cost of individual multi-rotor systems are illustrated in table 3.

Table 3. Individual price range for complete systems

Brand	Version	Price ($)
X3D	UFO	1,367
Draganfly	X4	8,495
	X6	19,999
	X8	32,165
Microdrone	Md4-1000	25,000
Mikrokopter	Basis L4	1,208
	Basis Hexa	1,654
	Basis Okto	2,068
X3D	UFO	1,367

3 Structural Performance

The structural performance is evidently one of the main factors of the design and this is where the majority of optimization can be achieved to reduce the overall mass by identifying the factor safety of individual components. The prototypes developed where tested using Finite Element Analysis (FEA) simulation before manufacture. These prototypes are then evaluated using an impact pendulum testing rig to simulate different velocities and impact energies and compare them with the FEA results. The mechanism itself proved to be successful but the surrounding shroud requires further development in order to guarantee repetitive quantitative results due to deflections.

4 Aerodynamic Performance

The aerodynamical effects of rotating blades were not taken into consideration in the simulation, as the main objective is to compare the amount of drag produced by the new additional components relative to the original existing frame. This provides early estimates of the additional drag, which when added to the additional mass will illustrate the overall reduction in achievable performance. The motor and blade combination used for the simulation is that of AXI 2826-14 and Epp. 1045 propellers.

The simulation was conducted at an overall laminar air flow of 8 m/s relative to the components.

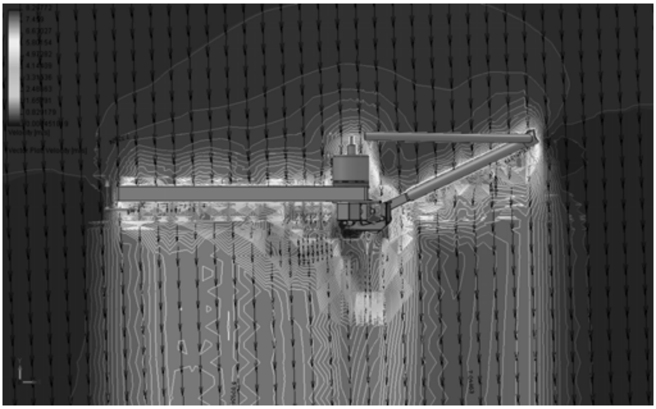


Fig. 4. Airflow simulation in SolidWorks. Light areas illustrate reduced air speed.

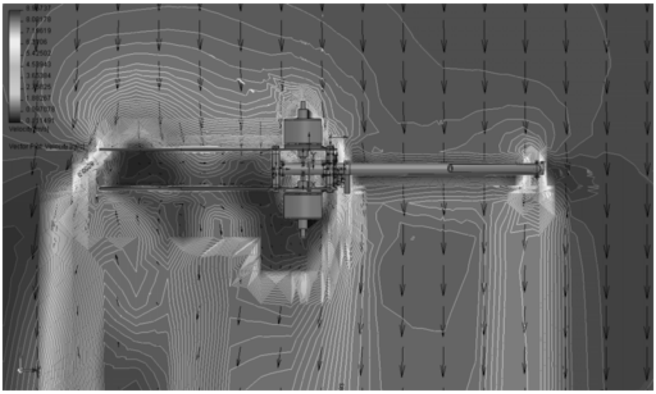


Fig. 5. Airflow simulation in SolidWorks. Light areas illustrate reduced air speed.

5 Conclusion

The system developed provides a benchmark from which further development can be achieved. Illustrated in this paper is a demonstration of a prototype which functions to its required ability, but is only at its first iteration. Further development can lead to other methods in the way the system functions. The performance requirements are rigorous or it is not feasible to have such a system attached to UAS. Areas to be further reviewed are the aerodynamical performance as well as the reduction in additional mass. A system such as this can provide for a greater variety of deployment scenarios, i.e. confined urban environments, capable of coping with inbound objects and out of range descent. The possibility of bringing this into a commercial context is being currently reviewed.

References

1. FAA, Aerospace Forecast Fiscal Report Years 2009 – 2025, p. 46, http://www.faa.gov
2. Hayward, K.: Unmanned Aerial Vehicles: A New Industrial System? http://www.raes.org.uk
3. United States Army Unmanned Aircraft Systems Roadmap 2010-2035: Eyes of the Army, http://handle.dtic.mil/100.2/ADA518437
4. CAA. Unmanned Aircraft System Operations in UK Airspace – Guidance, CAP 722
5. Kalinowski, N., Testimony – Statement of, FAA, July 15 (2010), http://www.faa.gov
6. Manning, S., Rash, C.E., LeDuc, P.A., Noback, R.K., McKeon, J.: The role of Human Causal Factors in U.S. Army Unmanned Aerial Vehicle Accidents. USAARL Report No. 2004-11
7. Williams, K.W.: A Summary of Unmanned Aircraft Accident/Incident Data: Human Factors Implications. FAA, DOT/FAA/AM-04/24
8. Hull, L.: Drone makes first UK ‘arrest’ as police catch car thief hiding under bushes, http://www.dailymail.co.uk
9. BAE Systems, Unmanned Air System Project For South Coast Formally Launched, Ref. 358, November 07 (2007), http://www.baesystems.com
10. Lewis, P.: CCTV in the sky: police plan to use military-style spy drones, January 23 (2010), http://www.guardian.co.uk
11. AR.Drone Parrot, http://ardrone.parrot.com
12. Prior, S.D., Shen, S.-T., White, A.S., Odedra, S., Karamanoglu, M., Erbil, M.A., Foran, T.: Development of a novel platform for greater situational awareness in the urban military terrain. In: Harris, D. (ed.) EPCE 2009. LNCS, vol. 5639, pp. 120–125. Springer, Heidelberg (2009)

Test-Retest Reliability of CogGauge: A Cognitive Assessment Tool for SpaceFlight

Matthew Johnston, Angela Carpenter, and Kelly Hale

Design Interactive Inc
1221 E Broadway, Suite 110
Oviedo, FL 32608
{Matthew,Angela,Kelly}@designinteractive.net

Abstract. The purpose of this study was to assess at a preliminary level, the test/retest reliability of the math processing mini-game of CogGauge: a cognitive assessment tool for spaceflight. The focus of this assessment was on the stability of test scores and calculation of reliable change on test/retest scores obtained on a mathematical processing task. A sample of 18 neurotypical, non-concussed individuals with a minimum of a graduate or professional school degree completed the task on two separate occasions separated by 7 days. Test-retest coefficients, reliable change difference scores (including adjustments for practice effects) and descriptive statistics are provided along with a discussion of the CogGauge tool.

Keywords: cognitive, decrement, assessment, diagnosis, reliability, stability.

1 Introduction

A cognitive assessment tool must be developed such that it is adequately sensitive to assess the presence of a decrement in cognitive functioning. The change in any score on the tool must be due to a real change due to said decrement and not an artifact of a lack of stability in the tool itself. Typically, this would require a tool to show stable performance over time when used by neurotypical or normal cognitive functioning individuals. For tracking the cognitive functioning of an individual, sessions could be spaced from six to 24 months but there are situations in which an individual may be assessed in much shorter increments. CogGauge is designed to provide a cognitive functioning baseline for an individual, to be used to track individuals over time (possibly 6-24 months) but also to be used when it is suspected that an individual has sustained a head injury or been exposed to an environment in which it is reasonable to assume their cognitive functioning is impaired. It is this unpredictability that may require the tool to be used in short intervals and therefore the analysis must account for potential practice effects.

Spaceflight presents a unique opportunity for the use of cognitive assessment tools in that an individual astronaut is exposed to both chronic and acute stressor potential such as microgravity and potential head trauma. Given that the duration astronauts spend in the space environment can have a wide range in duration, and the potential

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 565–571, 2011.

for unexpected acute trauma is present, the cognitive assessment analysis must account for potential improved performance and learning effects.

CogGauge (A Cognitive Assessment Tool for Spaceflight) is a computerized, game based cognitive assessment battery designed to assess cognitive functioning of astronauts. The test battery was inspired by validated neurocognitive assessment batteries such as The Automated Neuro-physiological Assessment Metrics (ANAM®; Reeves et al., 2007). Other cognitive assessment tools such as the Space Cognitive Assessment Tool (WinSCAT; Kane et al., 2005) have also used ANAM® battery elements. In the case of WinSCAT, ANAM® measures were adapted to meet NASA requirements and though it is sensitive to change in neurocognitive status of individuals due to head trauma (Levinson & Reeves, 1997) and fatigue induced by sleep deprivation among space crews (French et al., 1999), among other stressors, it is not a self motivating game based tool that attempts to engage astronauts in an entertaining experience. Alternatively, the 8 mini-games within the CogGauge test battery are designed to provide a more engaging experience that attempts to motivate astronauts through game based mechanisms but retains the test battery approach to maintain an efficient assessment procedure.

CogGauge consists of 8 mini games that assess short term memory, spatial processing and other perceptual abilities. CogGauge leveraged past research on the sensitivity of different neurocognitive tasks to stressors of interest during spaceflight to select the mini games for inclusion in the tool. Each game developed has shown sensitivity to at least one stressor that an astronaut may be exposed to during spaceflight. The specific stressors included are traumatic brain injury, radiation, sleep deprivation, physical fatigue, heat stress, microgravity, pressure and disease based cognitive deficits.

The purpose of this study is to assess the test-retest reliability, stability of performance of the math processing mini-game developed for CogGauge. This evaluation is preliminary and serves to support the iterative design process for CogGauge tool. This paper will review the test-retest reliability, practice effects and reliable change analysis results for a healthy population who completed the math processing task over a 7 day retest interval. Descriptive statistics are also provided.

2 Method

2.1 Participants

The participants were 18 healthy adults with no history of cognitive related disease or head trauma. No participants had been exposed to or were suffering the effects of the stressors of interest for spaceflight indicated previously within a minimum of 1 week prior to the test. There were 13 males and 5 females. Their average age was 31. All participants had obtained a graduate degree in an engineering or psychology field or had obtained a professional degree. Given the limited availability of astronauts, it was recommended by NASA to use degreed professionals that approximate the profile of potential astronaut recruits.

2.2 Procedure

Participants were asked to complete 20 trials of the mathematical processing on two occasions separated by 7 days. Attempts were made to ensure that participants completed the tasks at approximately the same time on each occasion.

2.3 Measures

CogGauge is a computerized game based cognitive assessment tool that consists of 8 separate mini games that measures attention, short term memory, response time and processing speed. For each mini game, accuracy, response time and throughput are calculated. In this paper, the mathematical processing mini game was evaluated.

Mathematical Processing – Asteroid Sling. This mini game is inspired by the mathematical processing task that is part of the ANAM® battery of tasks. The user is visually situated in the cockpit of a spaceship that is flying through an asteroid field. The user is presented with a mathematical equation on their heads up display as show in Figure 1.

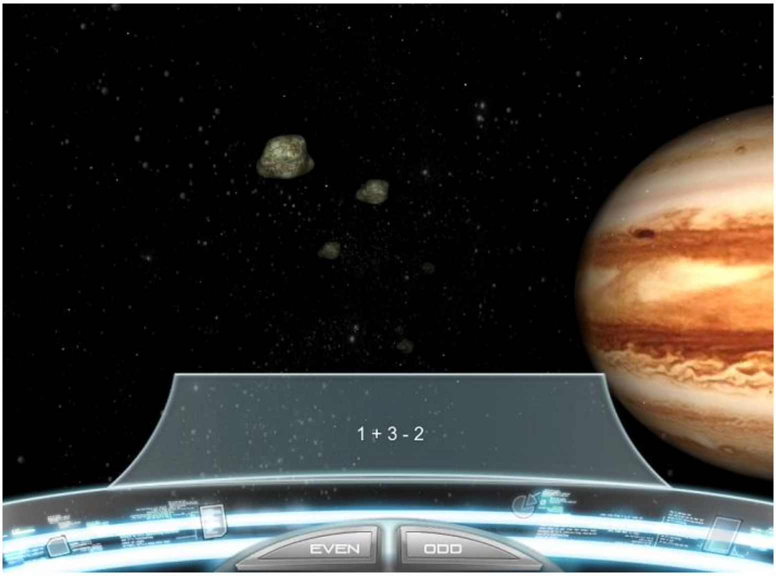


Fig. 1. Mathematical Processing Interface

The equation consists of three digits and two operators. Only single digits are used, and there is always an addition and subtraction operator. The addition operator and subtraction operator are presented as the first operator an equal number of times. At no point does the equation enable integer answers. The user is asked to answer the mathematical equation and determine if the answer is odd or even. They are then required to press the associated key (arrow keys on keyboard or left/right mouse buttons). A correct answer results in successful navigation of the asteroid field. An incorrect answer or not answering within the time limit (5000ms) results in a direct hit with an asteroid. The user performs 20 trials.

An accurate answer is defined as successful determination that the answer to the equation is an odd or even whole number and the response occurs within 5000ms. Response time is determined as the time from the presentation of the equation stimulus to the selection of the answer.

2.4 Analysis

This analysis was based on healthy, neurotypical individuals with no prior history or recent exposure to stressors that would be reasonably expected to result in a decrement in cognitive functioning. This was a within subjects design. Accuracy and average response time was calculated for each individual over the 20 trials on Day 1 and Day 2. Accuracy on each trial with was recorded as correct or incorrect. The percent correct for the 20 trials was stored for each user. Response time was averaged for the 20 trials and stored for each user.

Pearson correlation was used to determine the test-retest reliability of the math processing mini game for both accuracy and response time. The performance of the participants was examined using paired sample t-tests to determine if their performance improved significantly from day 1 to day 2.

Reliable change estimates were conducted two ways. The first was a method proposed by Jacobson and Truax (1991) which does not include modifications for practice effects.

$$RC = X_2\text{-}X_1 \,/\, S_{\text{diff}}\,. \tag{1}$$

where X1 represents a subject's baseline score, X_2 represents that same subject's assessment score, and S_{diff} is the standard error of difference between the two test scroes. S_{diff} can be computed from the standard error of measurement:

$$S_{\text{diff}} = \sqrt{}2(S_E)^2\,. \tag{2}$$

Standard Error of Measurement for a test, S_E, is in turn estimated from a test's standard deviation and reliability:

$$S_E = s_x\sqrt{}(1\text{-}r_{xx})\,. \tag{3}$$

where s_x is the standard deviation of the test-taker's scores and r_{xx} is the reliability of the test.

The second method accounts for practice effects (Parsons et al, 2009) in which improvements on the test from time 1 to time 2 across a similar population of users is incorporated into the analysis. This method provides an estimate of the expected improvement by an amount that is the approximate average of others from a similar population. Correcting for practice has been used previously in clinical psychology (Jacobson & Revenstorf, 1988; Speer & Greenbaum, 1995) and neuropsychology (Chelune, Naugel, Luders, Sedlak & Awad, 1993; Iverson, 2001) yet their remains no agreed upon method. In this evaluation the following equation was used:

$$\text{Practice-Corrected RCI} = ((X_2\text{-}X_1) - (M_2\text{-}M_1)) \,/\, \text{SDD}\,. \tag{4}$$

where X_1 is the subject's average baseline score, X_2 is the subject's average assessment score, M_1 is the group mean baseline score, M_2 is the group mean assessment (or post-condition) score, and SDD is the standard deviation of the group test-retest difference.

3 Results

Descriptive statistics and reliable change estimates for the mathematical processing mini game are shown in Table 1.

Table 1. Descriptive statistics for X neurotypical individuals

	M (SD) Test	*M (SD) Time Re-Test*	*p*
Response Time	2859.183 (632.01)	2701.989 (592.84)	.03
Accuracy	.875 (0.12)	.9056 (.078)	.055

The Pearson correlation coefficient for mathematical processing was 0.779 for accuracy and 0.835 for response time.

Group performance was compared using a paired sample t-test. There was a significant improvement from test to retest for response time (t(18), $p<0.05$). No difference was found for accuracy.

Participant level of performance was compared using a paired sample t-test within participants. 12 participants showed an improvement on response time from test to re-test sessions, 5 participants were slower. Two participants showed significant improvement from test to retest sessions ($p<0.05$). Seven showed an improvement on accuracy from test to re-test sessions while 3 decreased in accuracy. No changes in accuracy were significant. On average there was an average improvement in response time of 157ms and accuracy of 3% from test to re-test.

Reliable changes estimates revealed that only one participant showed reliable change in response time and accuracy (although different participants) from test to re-test. In both cases, this change was revealed using the calculation that corrects for practice only and a reliable improvement was found. No reliable decreases in performance were revealed.

4 Discussion

This study provides a preliminary evaluation of the test-retest reliability and stability of performance of two mini games within Cog Gauge, a cognitive assessment tool for spaceflight. Healthy, neurotypical participants were used in the evaluation to test the stability of the tool over a short interval. It is expected that in a short interval there would be some improvement on the test due to practice effects.

However, CogGauge was designed to reduce the effect of practice through randomization of stimuli presentation. Although the tool would be intended for use over longer intervals, such as 6 months, it is also intended to be used following exposure to a relevant stressor that could result in decrease cognitive functioning. For some stressors relevant to the astronaut domain, the presence of a stressor such as microgravity can be predicted in terms of onset of exposure. However, for other stressors such as traumatic brain injury the occurrence is unpredictable. Therefore the tool must be robust to use over short intervals.

The test-retest coefficients for the mathematical processing test were 0.779 and 0.835 for accuracy and response time respectively. These are comparable to the reported reliability of neuropsychological tests such as the Delis–Kaplan Executive Function System Trail- Making Test or Color-Word Test (Delis, Kaplan, & Kramer, 2001), the California Verbal Learning Test–Second Edition (Delis, Kramer, Kaplan, & Ober, 2000) and the Impact test (which reports scores between 0.65 and 0.8.

Group comparisons for response time were found to be significant, indicating the presence of practice effects on this particular measure when averaged across the group from baseline to the re-test. Although practice effects can be somewhat mitigated through randomization in the presentation of stimuli, there is still learning with respect to the interface, the understanding of instructions and the experience of being test. Given that a practice effect was discovered, it would suggest the use of the reliable change methodology that accounts for practice effects. However, there is no agreed upon criteria under which the method should be invoked. For instance, only one participant showed a significant different from test to re-test, however the group as a whole did show significant improvement. Should invoking this method be based on the individual or on the group performance improvement. How man individuals would need to show improvement, and by how much before the reliable change method for practice is invoked. Group comparisons did not reveal a significant improvement in accuracy but with a larger N it is expected it would have. Accuracy scores can have a direct relation to understanding instructions and becoming familiar with the entry of responses. Given that CogGauge attempts to be self administered and the interval from test to the retest was short, it is possible that participants did undergo familiarization. Again, this suggests that when short intervals are concerned and a participant is not practiced on the instrument, correcting for practice effects should be considered.

This study is limited in that it did not use the target user population, did not compare the short interval from test to retest to a longer interval and used a small sample. Future research should be conducted with a larger sample to further refine the interpretation of change on the full test battery. Unfortunately the target population, astronauts, is quite small with limited availability. Therefore a sample size that approximates the technical and mental abilities as well as potentially personality characteristics should be sought for future study. However, although CogGauge is targeted initially to astronauts, it could potentially be used by other populations such as warfighters, pilots or anyone exposed to stressors such as heat stress, sleep deprivation or physical fatigue among others. Providing users with example trials prior to full performance should also be considered to attempt to eliminate potential learning effects prior to performance.

5 Conclusion

The results of this evaluation indicate on a preliminary basis that the mathematical processing mini-game designed for inclusion in CogGauge: A cognitive assessment tool for spaceflight is reliably stable for response time and accuracy measures for a short test-retest interval. Practice effects are present when group data is evaluated, however evaluation at an individual level shows that improvement on average is not significant as only one participant showed significant improve on response time and accuracy scores. Future evaluations of this and other CogGauge mini games should investigate longer intervals for test-retest to understand stability and practice effects when a realistic assessment interval is applied, such as 6 months. When available, the sample population should be astronauts, although the sample size will most likely be small.

Acknowledgements. This material is based upon work supported b NASA under the SBIR program. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of or endorsement of NASA.

References

1. Chelune, G.J., Naugle, R.I., Luders, H., Sedlak, J., Awad, I.A.: Individual change after epilepsy surgery: Practice effects and base-rate information. Neuropsychology 7, 41–52 (1993)
2. Delis, D.C., Kaplan, E., Kramer, J.H.: Delis Kaplan Executive Function System technical manual. The Psychological Corporation, San Antonio (2001)
3. Delis, D.C., Kramer, J.H., Kaplan, E., Ober, B.A.: California Verbal Learning Test Adult Version manual, 2nd edn. The Psychological Corporation, San Antonio (2000)
4. Iverson, G.L., Green, P.: Measuring improvement or decline on the WAIS–R in inpatient psychiatry. Psychological Reports 89, 457–462 (2001)
5. Iverson, G.L., Lovell, M.R., Collins, M.W.: Interpreting Change on ImPACT following sports concussion. The Clinical Neuropsychologist 17(4), 460–467 (2003)
6. Jacobson, N.S., Revenstorf, D.: Statistics for assessing the clinical significance of psychotherapy issues: Issues, problems, and new developments. Behavioral Assessment 10, 133–145 (1988)
7. Jacobson, N.S., Truax, P.: Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology 59, 12–19 (1991)
8. Kane, R.L., Short, P., Sipes, W., Flynn, C.F.: Development and validation of the Spaceflight Cognitive Assessment Tool for Windows (WinSCAT). Aviation, Space, and Environmental Medicine 76 (Suppl. 6), B183–B191 (2005)
9. Parsons, T.D., Notebaert, A.J., Shields, E.W., Guskiewicz, K.M.: Application of Reliable Change Indices to Computerized Neuropsychological Measures of Concussion" International Journal of Neuroscience 119, 492-507 (2009)
10. Reeves, D.L., Winter, K.P., Bleiberg, J., Kane, R.L.: ANAM Genogram: Historical perspectives, description, and current endeavors. Archives of Clinical Neuropsychology 22(Suppl.1), 15–37 (2007)
11. Speer, D.C., Greenbaum, P.E.: Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. Journal of Consulting and Clinical Psychology 63, 1044–1048 (1995)

A Formalism for Assessing the Situation Awareness of Pilots

Steven J. Landry and Chittayong Surakitbanharn

School of Industrial Engineering, Purdue University
315 N. Grant St. West Lafayette, IN 47906, USA
slandry@purdue.edu, csurakit@purdue.edu

Abstract. The assessment of situation awareness is modeled within a set-theoretic formalism. This formalism explicitly requires the identification of a functional relationship between particular sets of knowledge and specific performance criteria. The framework is exercised in an experiment, demonstrating the utility of the formalism.

Keywords: situation awareness, aviation, simulation, set theory.

1 Introduction

The concept of situation awareness (SA) has been shown to be important to good performance [4], particularly in aviation [2]. While there have been some theoretical descriptions of SA (e.g., [1]; [3]; [6]), these have not been formalized. More specifically, all the main theories regarding SA are qualitative and have not achieved consensus from the academic community [5].

Formal definitions are useful for a number of reasons, including to ensure comparability between experimental results and to make the SA construct explicit. This paper presents such a formal framework for situation awareness, and discusses how it applies to various methods for assessing situation awareness. Specifically, we propose a set-theoretic framework for considering the concept of SA, which starts from only the assumption that SA is a set of identifiable knowledge available to the operator to support behavior. The remainder of the framework consists of logical inferences from this assumption. A general framework is constructed that has important implications for the concept and measurement of situation awareness.

The initial presentation of the framework provided in the next section is deliberately general. Despite the generality, the framework is intended to be mathematically rigorous, based primarily on set theory. Refinements, or possible refinements, to the framework are subsequently presented.

2 The Framework: Static Case

If we assume that certain knowledge is necessary to properly perform a task, then we can model the set of all such elements of knowledge as the "target set," as shown in

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 572–581, 2011.

(1). (This model is for the set of elements necessary at a given time t; the model is being extended to cover a dynamic SA.)

$$ {}_T\mathrm{K} = \left\{ {}_T\mathrm{K}_1, {}_T\mathrm{K}_2, \ldots, {}_T\mathrm{K}_n \right\} \tag{1} $$

Next, we define the "actual" set of knowledge available to the person(s), i.e. that person's situation awareness, as shown in (2).

$$ {}_{SA}\mathrm{K} = \left\{ {}_{SA}\mathrm{K}_1, {}_{SA}\mathrm{K}_2, \ldots, {}_{SA}\mathrm{K}_m \right\} \tag{2} $$

SA is either a subset of or equal to the person's total knowledge ${}_\Sigma\mathrm{K}$; i.e. ${}_{SA}\mathrm{K} \subseteq {}_\Sigma\mathrm{K}$. We next identify a "mapping function" f_{C_i} that relates a general set of knowledge to some performance criterion C_i. This mapping function relates the effect on the performance criterion of possessing K. (For example, if one of the elements of my SA is that my building is on fire, where the performance criterion is the probability I would evacuate, that probability would be higher than if I did not possess that knowledge.)

It is critical that the performance criterion be identified clearly; it cannot be assumed that the possession of a particular knowledge set will affect different performance criteria in the same way. Moreover, experiments must clearly tie experimental manipulations to the hypothesized relationship between a particular performance criterion and a knowledge set.

We then define subsets ${}_T^q\mathrm{K} \subset {}_T\mathrm{K}$ and ${}_T^r\mathrm{K} \subset {}_T\mathrm{K}$ such that:

$$ \exists C_i \text{ s.t.} f_{\mathrm{C}_i}\left({}_T\mathrm{K}\right) > f_{\mathrm{C}_i}\left({}_\neg\mathrm{K}\right), \forall {}_\neg\mathrm{K} \cap {}_T\mathrm{K} = \varnothing \tag{3} $$

$$ \exists {}_T^q\mathrm{K} \text{ s.t.} f_{\mathrm{C}_i}\left({}_T^q\mathrm{K}\right) \geq f_{\mathrm{C}_i}\left({}_T^r\mathrm{K}\right), \forall r \neq q \tag{4} $$

where:

$$ f_{\mathrm{C}_i}\left(\mathrm{K}\right) = C_i \tag{5} $$

Equation (3) states that there exists some performance criterion C_i such that performance given the knowledge from the target set is better than if no knowledge from the target set is available. Equation (4) states that there exists some subset of the target set that results in as good as, if not better, performance than any other subset.

We argue that unless (3) is true, situation awareness has no value as a construct. That is, if it is possible to achieve all performance criteria without the need to possess any separately identifiable set of knowledge, situation awareness is not a useful concept. (It is possible that such a set is non-exclusive; there may be multiple sets, where any one of those sets is capable of providing the best performance; it is only necessary that these sets provide better performance than other sets.)

An implication of equation (3) is that only those C_i where (3) is true represent performance criteria affected by situation awareness. This is critical, as it implies that there are performance criteria for which the concept of situation awareness is not useful; in such cases, we cannot identify a particular set of knowledge that results in better performance with respect to those criteria than any other set of knowledge. This result allows us to identify how SA differs from other concepts.

2.1 Example: Static Case

The criterion C_i can be defined in a variety of ways, including as executing a particular behavior, accomplishing a task in each individual instance, the probability of accomplishing a task over repeated trials of the task, achieving of a certain level of task quality, or completing the task within a particular period of time. There can be multiple criteria for a particular task. (Again, although C_i can be defined in many ways, it must be defined specifically; the mapping function, however, from ${}_{T}\mathrm{K}$ to C_i does not need to be identified but can be determined empirically.)

For example, one might define the following two criteria for a task:

$$C_1 = \text{Push stop button before } t = 45 \text{ seconds}$$
$$C_2 = \text{Mean accuracy}$$

One would then identify a mapping function relating certain knowledge to these performance criteria:

$$f_{C_1}(\mathrm{K}) = \begin{cases} 1, \text{"Fire exists"} \in \mathrm{K} \\ 0, \text{"Fire exists"} \notin \mathrm{K} \end{cases}$$

$$f_{C_2}(\mathrm{K}) = \begin{cases} \alpha, \text{"Wind direction","Wind speed","Object speed"} \in \mathrm{K} \\ \alpha - \varepsilon_1, \text{"Wind direction"} \notin \mathrm{K} \\ \alpha - \varepsilon_2, \text{"Wind speed"} \notin \mathrm{K} \\ \alpha - \varepsilon_3, \text{"Object speed"} \notin \mathrm{K} \end{cases} \qquad .\alpha, \varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$$

We identify ${}_{T}\mathrm{K} = \{\text{"Fire exists","Wind direction","Wind speed","Object speed"}\}$. We then note that any set of knowledge ${}_{\neg}\mathrm{K}$ not containing any of these elements, i.e. where ${}_{\neg}\mathrm{K} \cap {}_{T}\mathrm{K} = \varnothing$ would result in worse performance, i.e. $f_{C_i}\left({}_{T}\mathrm{K}\right) > f_{C_i}\left({}_{\neg}\mathrm{K}\right)$. Equation (3) is then true; SA is a useful construct for these performance criteria. Also, ${}_{T}^{1}\mathrm{K} = \{\text{"Fire exists","Wind direction","Wind speed","Object speed"}\}$ results in better performance than any other subset of ${}_{T}\mathrm{K}$. In addition, note that sets where ${}_{T}\mathrm{K} \subset {}_{\Sigma}\mathrm{K}$ all result in as good a performance as with ${}_{T}\mathrm{K}$, so the set is non-exclusive.

3 Mapping of Existing Measurement Methods to the Framework

A number of methods have been used to measure SA, including “direct” methods, such as the Situation Awareness Global Assessment Technique (SAGAT), the Situation Present Assessment Method (SPAM), (SALSA), “indirect” methods, and subjective methods such as Situation Awareness Rating Technique (SART). These methods carry certain assumptions that can be critically examined within the framework. Each will be discussed separately in the following sections.

3.1 SAGAT

SAGAT consists of having an operator perform a task in simulation which is periodically stopped for situation awareness queries. During the stops, access to information displays is removed, and the operator is asked a number of questions. The duration of the stops is generally limited to less than two minutes, with the target number of questions being 7 within that time frame. (There is substantial latitude in how the procedure is applied.) The specific queries are randomly selected from a set of relevant information gleaned by a goal directed task analysis. The accuracy of the operator at answering the questions is a reflection of that person's "global" situation awareness.

The results of the goal directed task analysis is ${}_{T}\mathrm{K}$. SAGAT then randomly or comprehensively chooses elements of ${}_{T}\mathrm{K}$ and queries the operator to determine if those elements are part of the operators' ${}_{SA}\mathrm{K}$. The results of the queries are used to estimate the operator's global situation awareness, essentially estimating the ratio of the cardinalities of the sets, i.e. $\left|{}_{SA}K\right| / \left|{}_{T}\mathrm{K}\right|$, by the quotient of the quantity of correct answers and the total number of questions. (For example, if the participant got 6 questions right out of 10, the estimate is 0.6.) Sometimes the questions are questions that can be strictly graded as correct or incorrect; other times the rater must establish a criteria for considering the question correct or incorrect. The set of questions can be broken into categories, such as the "level" of situation awareness.

This score is then argued to correlate with overall performance. In general, a stronger conclusion is asserted based on past validation studies – that higher SA results in better overall performance. Considered within the proposed framework, there is only one performance criterion in question, "overall performance." The implied function is, roughly:

$$f_C(\mathrm{K}) = \begin{cases} \alpha, {}_{T}\mathrm{K} \subset \mathrm{K} \\ \alpha - \Delta, {}_{T}\mathrm{K} \not\subset \mathrm{K} \end{cases} \quad \text{where } \Delta = f\left({}_{T}\mathrm{K} \cap \mathrm{K}\right)$$

It is argued that this view of SAGAT is consistent with the method. However, making the method explicit yields important insights. First, the method frequently utilizes random sampling to estimate the measure, but sampling error, as quantified in the standard error of the estimator, is never identified. It is therefore difficult to know the accuracy of the conclusions drawn through the use of SAGAT.

Second, the values of α and, more importantly, Δ are never estimated, and may vary widely across application. Moreover, Δ is an undefined function of the intersection between the person's knowledge and the target set. So even if one agrees in principle with the claim that better SA results in better performance, the performance benefit from a given SA set is unknown.

Within the framework, one cannot know whether equation (3) and (4) are satisfied unless this function is identified. It is assumed that this function takes on positive, non-zero values. In that case, SAGAT's results would support equation (3) and (4) and indicate that the concept has value.

3.2 SPAM

SPAM was introduced to address criticisms that SAGAT was focused on knowledge in memory, whereas it has been argued that operators might offload some knowledge to the environment. SPAM introduces queries without stopping the simulation, leaving the environmental information available to the operator during the queries. Because the information is available, accuracy is presumed to not be a good measure of SA; instead, time to respond is used. The argument is that if the environmental information is part of the SA of the operator, then the operator will respond more quickly to the queries when controlled for workload.

Presumably, ${}_{T}\mathrm{K}$ can be identified similarly to SAGAT, and a random sample of ${}_{T}\mathrm{K}$ can be queried. Time to respond to queries, after indicating that the operator is ready to answer a query, is then used as the measure. Sometimes, incorrect answers are eliminated from the scoring, although again the method is not specific in this regard. Within the framework, the mapping function is then:

$$f_C(\mathrm{K}) = \alpha + \beta\big(RT(\mathrm{K})\big) \quad \beta < 0 \tag{6}$$

where:

$$RT(\mathrm{K}) = \begin{cases} \alpha_i, {}_{T}\mathrm{K} \subset {}_{SA}\mathrm{K} \\ \alpha_i + \Delta_i, {}_{T}\mathrm{K} \not\subset {}_{SA}\mathrm{K} \end{cases} \quad \forall i, \Delta = f\left({}_{T}\mathrm{K} \cap \mathrm{K}\right)$$

Equation (6) states that overall performance is inversely proportional to the response time of a person to the SPAM queries; that response time is lower if the target set of knowledge is part of the person's SA. (The subscript *i* refers to the participant; it is not presumed as part of SPAM that the response time, or the difference in response time, is constant across individuals.) Presumably, this relationship holds whether the information is in memory or in the environment.

Given (6), one can conduct tests under different conditions, with the condition yielding the faster response times concluded to produce better situation awareness. That is, SPAM assumes that $RT(\mathrm{K}) < RT(\mathrm{K}^*) \rightarrow \mathrm{K} \cap {}_{T}\mathrm{K} > \mathrm{K}^* \cap {}_{T}\mathrm{K}$ and $\mathrm{K} \cap {}_{T}\mathrm{K} > \mathrm{K}^* \cap {}_{T}\mathrm{K} \rightarrow f_C(\mathrm{K}) > f_C(\mathrm{K}^*)$. However, there is no proof of these assertions, nor is there any convincing empirical support. (This is not to say these assertions are incorrect or unreasonable.)

An alternative is that SPAM defines SA as consisting of items that can be accessed in less time than other items. However, what constitutes "less," and whether there are two distinct classes of items, in SA and not in SA, or the items fall along a continuum, is not specified.

3.3 SALSA

SALSA was developed to address variability in the importance of items, under the assumption that operators may be more likely to have important items in SA when compared to less important items. This method implies that an SA score derived from a method such as SAGAT may not be ordinal; that is, a higher SA score may not be better if that SA is composed of unimportant items. SALSA therefore argues that one must concentrate SAGAT queries on important items. Specifically, SALSA, as applied to air traffic control, includes a prior step that identifies the relative importance of aircraft; SA queries are then focused on these aircraft only.

SALSA is very similar to SAGAT, with the exception that the sample is not randomly selected from ${}_{T}\mathrm{K}$. Instead, a subset of ${}_{T}\mathrm{K}$, presumably composed of those items that have a substantial effect on performance, are sampled. The effect of this manipulation on the standard error of the estimator is unknown.

3.4 Indirect Methods

Indirect methods consist of those that infer SA from performance. That is, rather than query the SA of the person, where performance is the mapped, performance is directly measured, and SA is mapped from that performance. Within the framework, C_i is measured, and the "inverse function" is used to calculate ${}_{SA}\mathrm{K}$. Symbolically:

$$\mathrm{K} = f_{C_i}^{-1}\left(C_i\right) \tag{7}$$

In the example above, $C_1 = \text{Push stop button before } t = 45 \text{ seconds}$. Applying (7) to that example, if $C_1 = 1$, we can infer $\text{"Fire exists"} \in \mathrm{K}$, because the mapping function is binomial. (That is, if $C_1 = 0$, then $\text{"Fire exists"} \notin \mathrm{K}$; these are the only two possibilities for C_1 .) However, in cases where the performance criteria is not so clearly mapped, such as for C_2 , then SA might not be able to be inferred from indirect measures.

Indirect methods are probably not capable of producing a "global" SA measure, such as that produced by SAGAT. Without an identification of the Δ function between having particular elements in SA, it is not possible to apply the inverse function to obtain the elements of the operator's SA set.

However, for the particular cases where an inverse function exists and is known, indirect methods can be highly effective. Such cases are probably limited to specific, and not global, performance criteria.

3.5 Subjective Methods

Subjective methods still posit that a relationship exists between SA and performance, but do not attempt to measure SA or performance. Instead, a subjective assessment is made of the person's SA by an evaluator. (Typically the evaluator is a subject matter expert.) This formalism does not address subjective methods.

4 Experiment

A flight simulator experiment was designed to exercise this framework. Each flight profile was designed to have the participant encounter a focus event, for which performance on the focus event could be evaluated. In each case, the information required for proper performance was available to the participants during the scenario. Specifically, the scenario descriptions and focus events are shown in Table 1.

The purpose of the experiment was, in part, to test our ability to validate or invalidate such mapping functions. That is, this experiment was primarily designed to exercise the framework rather than to test the particular relationships identified. As such, of interest was simply the ability to posit a relationship between specific set of knowledge and performance, and then to design an experiment to test that relationship. (Whether the relationship was true or false was unimportant.) Two mappings were used – one using SAGAT and one using the indirect method.

Table 1. Scenario descriptions and focus events

Scenario description	Focus event
Aircraft is executing an instrument landing system approach and landing.	Aircraft is aligned to land on runway 25 but winds favor runway 7.
Aircraft is descending from 8,000 ft. to start an approach.	Aircraft encounters icing as indicated by unusual instrument readings.
Aircraft descends from 8,000 ft. and begins an approach.	Aircraft is given vectors that would cause a near mid air collision.
Aircraft is in cruise at 4,000 feet and is instructed to descend to 3,500 feet.	At 3,500 feet, the aircraft would be below the minimum reception altitude for the airway.
Aircraft is at 6,000 feet, approaching the airport.	A vacuum system failure occurs.

4.1 Participants

Eighteen (18) pilot participants flew five flight simulator profiles after completing one training scenario in order to become accustomed to the controls and testing methods. The duration of the scenarios was between 15 and 30 minutes each. All but two of the pilots were between 18 and 26 years old. (One was under 50 and another was over 50.) Six (6) of the pilots had over 500 total flight hours; the remainder had between 200 and 500 hours.

4.2 SAGAT Mapping

Two types of SAGAT questions were used. First, general SAGAT questions were selected at random from general state information needed for good pilot performance. (For example, pilots were asked about altitude, heading, orientation, speed, and position of nearby aircraft.) A second set was related to the focus events, but were indirect. For example, in scenario 1, where the focus event was whether the pilot would land opposite the direction indicated by the wind, pilots were asked for the wind direction at the airport. (They were not asked whether they were landing opposite the wind.) The mapping implied by SAGAT, as indicated in section 2.2, is general, i.e. that:

$$f_C(\mathrm{K}) = \begin{cases} \alpha, {}_T\mathrm{K} \subset \mathrm{K} \\ \alpha - \Delta, {}_T\mathrm{K} \not\subset \mathrm{K} \end{cases} \quad \text{where } \Delta = f\left({}_T\mathrm{K} \cap \mathrm{K}\right)$$

The SAGAT score (*S*) is an estimate of ${}_{T}\mathrm{K} \cap \mathrm{K}$. From this, the hypothesis is then: $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$. (This hypothesis will be checked by conducting a binary logistic regression to see if the slope of a line fitting correct performance against SAGAT score.) The degree of this relationship, specifically the magnitude of Δ, can be estimated from this exploratory study and used as a specific hypothesis for future focused studies.

4.3 Indirect Method

The inverse mappings proposed for the scenarios are as follows:

1. $$\mathrm{K} = f_{C_i}^{-1}(C_i) = \begin{cases} \text{Winds favor runway 7} \in {}_{SA}\mathrm{K}, C_i = \text{participant lands on runway 7} \\ \text{Winds favor runway 7} \notin {}_{SA}\mathrm{K}, C_i = \text{participant lands on runway 25} \end{cases}$$

2. $$\mathrm{K} = f_{C_i}^{-1}(C_i) = \begin{cases} \text{Icing is present} \in {}_{SA}\mathrm{K}, C_i = \text{anti-icing equipment turned on} \\ \text{Icing is present} \notin {}_{SA}\mathrm{K}, C_i = \text{anti-icing equipment remains off} \end{cases}$$

3. $$\mathrm{K} = f_{C_i}^{-1}(C_i) = \begin{cases} \text{Aware of conflicting aircraft} \in {}_{SA}\mathrm{K}, C_i = \text{participant turns} \\ \text{Aware of conflicting aircraft} \notin {}_{SA}\mathrm{K}, C_i = \text{participant does not turn} \end{cases}$$

4. $$\mathrm{K} = f_{C_i}^{-1}(C_i) = \begin{cases} \text{Aware of MRA} \in {}_{SA}\mathrm{K}, C_i = \text{adheres to MRA} \\ \text{Aware of MRA} \notin {}_{SA}\mathrm{K}, C_i = \text{remains at 3,500 ft.} \end{cases}$$

5. $$\mathrm{K} = f_{C_i}^{-1}(C_i) = \begin{cases} \text{Vacuum failure} \in {}_{SA}\mathrm{K}, C_i = \text{participant lands on runway 7} \\ \text{Vacuum failure} \notin {}_{SA}\mathrm{K}, C_i = \text{participant lands on runway 25} \end{cases}$$

It is hypothesized that without the requisite information, good performance should be almost impossible, and with the requisite information, good performance should be almost certain. To ensure the results are accurate, post-scenario debriefs were conducted as confirmation.

4.4 Dependent Variables

For half of the trials, SAGAT probes were performed during the scenarios. The other half served as a control so that the effect of asking the SAGAT questions, if any, could be identified. The presence of SAGAT (or not) was counter-balanced across participants.

Specifically, for the SAGAT trials, 6 SAGAT questions were asked after stopping the scenarios, blanking the screens, and turning the participants away from the screens. For each participant, this was done twice during each SAGAT scenario. (About ½ of each participant's trials were SAGAT trials.) The questions were presented in written form and the participants wrote down their answers. Two of these questions were designed to query information relevant to the focus event; the other questions were designed to query general SA.

Performance of the participant for the focus event was evaluated as either correct or incorrect based on criteria established prior to the experiment. This focus event performance data was evaluated for this paper. General performance data, including complete aircraft state and control information, was sampled at 1Hz. Evaluation of general performance data is not included in this paper.

5 Results

There were 45 branching event responses each for both SAGAT and non-SAGAT trials. For the SAGAT trials, participants performed the focus event correctly 40% of the time. For the non-SAGAT trials, participants performed the focus event correctly 51% of the time. According to a two-sample test of proportion, this effect was not statistically significant ($p > 0.2$).

5.1 SAGAT

To determine if performance of the branching event was better for subjects with better SAGAT scores, two binary logistic regressions were performed. One fit performance against the overall SAGAT scores; the other fit performance against the SAGAT scores on the questions related to the branching event.

For the former, the regression was significant, p =0.032, G = 4.591. The odds ratio for the coefficient of SAGAT scores was 76.82, the log-likelihood ratio for the regression was -30.91, and 58.6% of the pairs were concordant, resulting in a Somer's D of 0.32 and a Goodman-Kruskal γ of 0.37. The regression equation for this case is z = -3.76 + 4.34 (SAGAT score), where SAGAT score is normalized as a percentage of the 12 questions answered correctly.

A second regression was performed using just the proportion of the branching-event related questions answered correctly. That regression was significant, p =0.027, G = 4.891. The odds ratio for the coefficient of SAGAT scores was 16.41, the log-likelihood ratio for the regression was -30.76, and 52.6% of the pairs were concordant, resulting in a Somer's D of 0.33 and a Goodman-Kruskal γ of 0.45. The regression equation for this case is z = -2.52 + 2.80 (SAGAT score on branching event questions).

5.2 Inverse Mappings

The inverse mapping suggests that performance can be used to infer knowledge. For the cases where performance of the branching event was correct, knowledge should have been nearly perfect. A 95% confidence interval on the median proportion of correct branching event questions was {0.75, 1.0}; in 8 out of 19 cases of good branching event performance participants got all branching event questions correct. In cases where the branching event was not performed correctly, 7 out of 31 participants answered all the branching event questions correctly. While the difference in the proportions was significant, $t = -2.24$, $p = 0.031$, the inverse mapping did not appear to be strongly related to knowledge in the way implied by the hypothesized relationship.

Post hoc elicitation was used to evaluate the veracity of the inverse mappings. That is, subjects were queried after the scenarios about their knowledge of the focus event. In twelve (12) cases, subjects indicated they were aware of the focus event, even though they responded to it improperly. The reasons for not responding to the focus event included "simulator effects," where the subject assumed the event was unintentional and ignored it, and a more complex consideration of the event than was considered by the experimenters. (For example, one pilot chose to land with the tailwind in scenario 1 because "the runway was long enough for (a) tailwind landing."

6 Discussion

Asking SAGAT questions resulted in an 11% improvement in performance at the focus events. Although this was not statistically significant, a 10% difference may be practically significant and should be investigated with a larger sample size and a more focused experiment. The logistic regressions suggest that SAGAT, particularly branching event-related questions, were predictive of good performance at the branching event, as hypothesized. The magnitude of the effect () appeared to be on the order of a 10% higher likelihood of success for each additional 25% of the branching event questions correct. This, however, is a very rough and probably task-specific result. The SAGAT mapping, however, appears to have some merit, sufficient to continue experimentation along this line. However, the inverse mappings, which were carefully selected prior to the experiment, were not sufficiently predictive. This casts doubt on the utility of indirect methods for assessing situation awareness, as such mappings appear highly sensitive to task environment effects (at least).

Acknowledgments. This work was supported by NASA Ames Research Center under cooperative agreement number NNA06CN30A. Walt Johnson was the technical monitor.

References

[1] Banbury, S., Tremblay, S.: A cognitive approach to situation awareness: theory and application. Ashgate Pub. Ltd (2004)

[2] Durso, F.T., Sethumadhavan, A.: Situation awareness: Understanding dynamic environments. Human Factors 50, 442–448 (2008)

[3] Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors 37, 32–64 (1995)

[4] Parasuraman, R., Sheridan, T.B., Wickens, C.D.: Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. Journal of Cognitive Engineering and Decision Making 2(2), 140–160 (2008)

[5] Stanton, N.A., Salmon, P.M., Walker, G.H., Jenkins, D.P.: Is situation awareness all in the mind? Theoretical Issues in Ergonomics Science 11, 29–40 (2010)

[6] Stanton, N.A., Stewart, R., Harris, D., Houghton, R.J., Baber, C., McMaster, R., et al.: Distributed situation awareness in dynamic systems: Theoretical development and application of an ergonomics methodology. Ergonomics 49, 1288–1311 (2006)

Mental Resource Demands Prediction as a Key Element for Future Assistant Systems in Military Helicopters

Felix Maiwald and Axel Schulte

Universität der Bundeswehr München (UBM), Department of Aerospace Engineering,
Institute of Flight Systems (LRT-13), 85577 Neubiberg, Germany
{felix.maiwald,axel.schulte}@unibw.de

Abstract. This work presents an approach to enhance knowledge-based assistant systems in the domain of military helicopter missions with the ability to prevent the pilot from being overtaxed. Therefore, an estimation method for residual mental capacity and current subjective workload is proposed. This estimation enables the assistant system to deduce the pilots' specific needs of support. As a result the assistant system shall be enabled to cooperate with the pilot by resource adaptive information exchange. First evaluation experiments of the prototype, conducted in our research helicopter mission simulator, will be described.

Keywords: Task load, subjective workload, mental resources, adaptive automation.

1 Introduction

Today's army helicopter missions are hardly imaginable without the support of Unmanned Aerial Vehicles (UAVs). In order to facilitate the gathering of closed-loop near real-time tactical reconnaissance data, future UAVs will not only be controlled from ground stations, but also guided from aboard a manned air vehicle. The on-board control of UAVs will add a new and broader range of tasks for the cockpit crew – for the commander as well as for the pilot flying. Due to the extra work demands put on the commander, one critical consequence certainly is the reassignment of cockpit tasks from the commander to the pilot flying, which could result in a noticeable increase of subjective workload. This new spectrum of tasks for the pilot flying, mainly in radio-communication, routing during flight and ad-hoc mission re-routing, is comparable to a single pilot environment.

In this context, the MiRA (Military Rotorcraft Associate) knowledge based assistant system for pilot flying has been developed and tested at the UBM. MiRA and other state-of-the-art approaches of knowledge based assistant systems (e.g. CAMA [8], PA [2]) focus on the provision of support functions in certain predetermined situations. However, they usually detect currently occurring violations of flight and mission parameters and use these as triggers to provide support. The assistant system corrects these errors and prevents further consequences. Suchlike reactive assistant system behavior may result in an overload for the pilot, especially if

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 582–591, 2011.

the system intervenes at a time when the human operator has no more free cognitive resources to adopt the offered support or interact with the system.

To prevent such situations, within this article an enhancement for the MiRA-assistant system is investigated, which shall allow a model based prediction of the human operator's workload and the remaining capacity of his/her mental resources in the current task situation.

2 Model Based Human Resource and Workload Determination

The Institute of Flight Systems of UBM investigates so called MUM-T (manned-unmanned teaming) approaches for military helicopter missions. Due to the expected reassignment of tasks from the commander to the pilot, we offer support to the pilot by knowledge based assistant systems. This enables the pilot to cope with an extended task load spectrum. The requirements of such an assistant system are described below.

To investigate the interactions between pilot, assistant system and the helicopter, the approach of the work system (introduced by [7]) will provide a framework. A work system refers to the components that perform the work process for achieving a given work objective. This consideration of the work system focuses on the physical dimension of the process on the participating entities.

In contrast to recent approaches of assistant systems (e.g. RPA [5]), we claim an Artificial Cognitive Unit (ACU) to assist the pilot as a co-operative automation element. This mode of automation is characterized by the fact that both the pilot and ACU determine and supervise what will happen in the course of the work process and which Operation-Supporting Means (OSMs) will be deployed at what time.

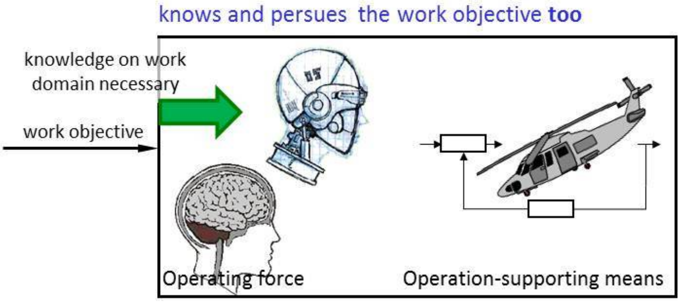

knows and persues the work objective too
knowledge on work domain necessary
work objective
Operating force
Operation-supporting means

Fig. 1. Assistant system as part of the Operating Force

Figure 1 depicts the first characteristic of such an ACU, being a part of the Operating Force (OF). Hence, not only the human pilot but also the ACU needs to understand the given work objective. Pursuing a common objective is only possible, if the assistant system is able to understand the necessary tasks to be performed. So, task models have to be developed, which incorporate the a-priori knowledge on military transport helicopter missions. An example for task models on military transport missions for aircrafts can be found in [9]. Secondly, we demand the ACU to co-operate with the pilot as an assistant system like it would be the case if there was a human assistant (see Figure 2). Hence, it takes initiative on its own in order to accomplish a given common work objective.

Swezey & Salas [11] describe desired capabilities of team members to ease interaction for the sake of team performance. The most important capability, pointed out for designing this work system, is that "effective team members typically help other team members who are having difficulty with a task". Conversely this means, overtaxing of a team member should be avoided.

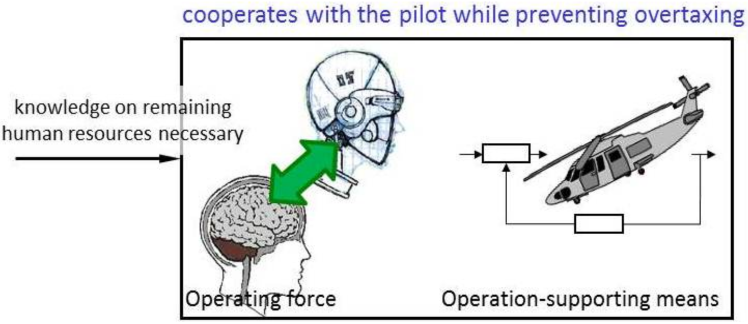

cooperates with the pilot while preventing overtaxing
knowledge on remaining human resources necessary
Operating force
Operation-supporting means

Fig. 2. Assistant system Co-operating with the human Operator

Veltman and Jansen [12] describe an inverted U-shaped relationship between the subjective workload (coming along with the work demands, i.e. the task load) and the human work performance (see Fig. 3). At a moderate workload level, performance can be increased to a peak plateau (optimal performance). In situations of high workload, the achievable performance decreases again.

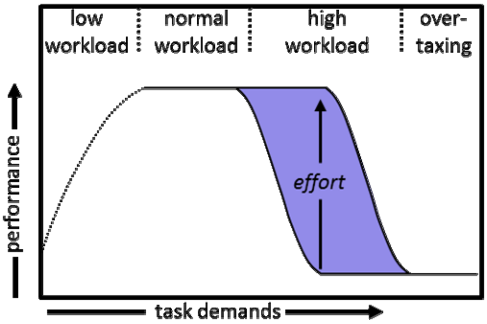


Fig. 3. Relationship between task load, workload and performance [12]

For the reasons mentioned above an ACU should be designed to keep the human workload, and therefore the task load on a moderate level, which permits the pilot to achieve best performance. In order to do so, the ACU has to estimate the demand on human resources. This estimation shall be done taking into consideration the current task situation. A similar approach of workload determination has been worked on in in the RPA program [5]. We particularly investigate an approach considering multiple task situations.

Considering human-machine co-operation, an ACU should be able to make use of the OSMs (see figure 4). To prevent the pilot from overtaxing (see Figure 2), we demand the ACU to adapt all dialogs with regard to the remaining resources as well as to the current workload of the pilot. For this purpose, the ACU has to choose appropriate human modalities to push any information to the pilot for fulfilling the goal of resource-oriented information transfer.

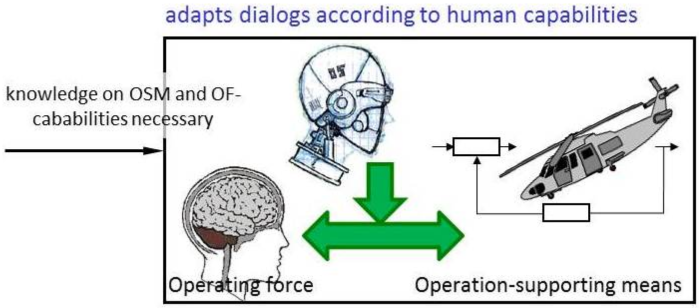

adapts dialogs according to human capabilities
knowledge on OSM and OF-cababilities necessary
Operating force
Operation-supporting means

Fig. 4. ACU has to choose appropriate modalities in order to transfer information

3 Detailed Concept

The concept described in the following section and shown in Figure 5 is based on Wickens' model of human cognitive abilities in information acquisition, processing and response [13].

In the first step, an ACU designed according to this approach has to capture as many as possible external information relevant to the pilot during a military transport helicopter mission (i.e. the state of the helicopter, the mission objective, the current flight and mission phase as well as environmental conditions).

After aggregation of this data into a comprehensive situational picture, this information can be used to determine the current tasks, the pilot should be executing momentarily and in the near future. For this purpose models of mission-typical task situations have to be developed on the basis of knowledge acquisition experiments. Such task models represent the rule-based normative behavior of the pilot, as it would be in accordance with aviation regulations and flight manuals [9].

To synchronize the assumed tasks with the tasks the pilot is actually executing, human machine interactions such as visual information acquisition (e.g. gazing at moving map) as well as manual interactions shall be analyzed. Therefore, assumptions are made that observing the gazes [3, 10] as well as observing the manual interactions enables the assistant system to draw conclusions on the tasks actually processed by the human operator.

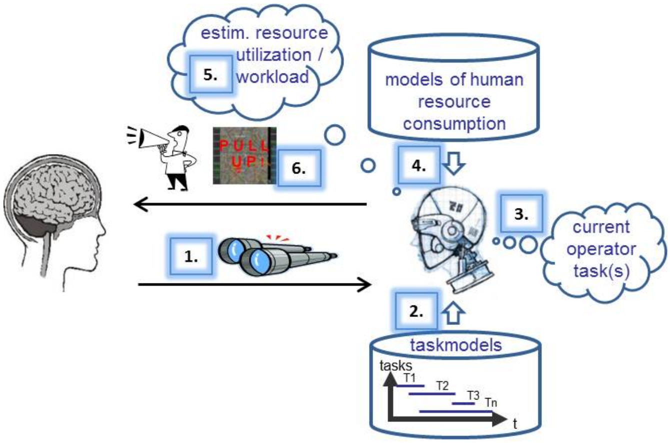

estim. resource utilization / workload
5.
models of human resource consumption
PULL UP!
6.
4.
3.
current operator task(s)
1.
2.
taskmodels
tasks
T1
T2
T3
Tn
t

Fig. 5. Concept for estimating remaining human resources / current workload

In a further step the determined actual task(s) shall be associated with the required human resources, by use of eight-dimensional demand vectors [6].

According to Wickens' [13] so called multiple-resource theory, every demand vector symbolizes the demand a single task poses on the human operator expressed in the terms of information acquisition, information processing and response. To explain the demand vector in detail, a sample for the task "navigate to next waypoint" is illustrated below in Figure 6. To accomplish this task, the pilot has to acquire visual spatial information from the map and the outside view, process this information by the use of cognitive-spatial resources whether he is on track and at last he has to decide and execute a manual control interaction.

Task	demand vector							
	perception				cognition		response	
	visual-spatial	visual-verbal	auditiv spatial	auditiv verbal	cog.-spatial	cogn.-verbal	man. control	verbal
Navigate to next waypoint	1	-	-	-	1	-	1	-

Fig. 6. Demand vector for the sample task "navigation"

In order to improve the sensitivity of successive resource and workload estimation, all ratings of demand vectors shall be classified according to the following task levels.

- sub-conscious tasks (manual control)
- routine tasks (supervisory control)
- tasks requiring planning and problem-solving

This classification shall provide a basis for a modification of each demand vector.

For the estimation of resource utilization, a Visual-Auditory-Cognitive-Psychomotor (VACP) model [1] shall be customized to make it use for the eight-dimensional demand vectors as input values. This enables the assistant system to determine the remaining resources and can be used to properly select a free resource for interaction with the pilot.

The W/INDEX-model [6, 14] shall be used as a basis for the estimation of the current workload of the pilot. Due to the limitation of W/INDEX-model handling two tasks at once only, the original method of W/INDEX has to be enhanced for multiple task situations. So the modified W/INDEX model has no limitation on the number of tasks to be processed in parallel [4].

The workload estimation enables the assistant system to decide whether a counteraction would be necessary. A simple way of workload reduction can be realized e.g. via a delay of low prioritized interactions with the pilot.

Moreover the resource consumption model provides an estimation, which human mental resource for information acquisition and processing would be the most suitable one for any necessary information transfer through the assistant system in the current situation. For this purpose possible future assistant system initiatives shall be analyzed using the modified W/INDEX model in order to identify human resource conflicts with respect to the present tasks. As a consequence this estimation enables the ACU to charge human resources equally while generating the lowest cost for the pilots' workload.

4 Module Architecture of Laboratory Prototype

To realize the presented concept, the following module structure (see Figure 7) has been sketched and followed through the development of a laboratory prototype.

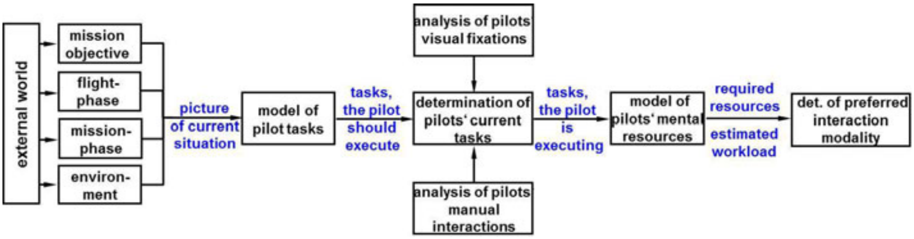

external world
mission objective
flight-phase
mission-phase
environ-ment
picture of current situation
model of pilot tasks
tasks, the pilot should execute
analysis of pilots visual fixations
determination of pilots' current tasks
analysis of pilots manual interactions
tasks, the pilot is executing
model of pilots' mental resources
required resources
estimated workload
det. of preferred interaction modality

Fig. 7. Module architecture of resources / workload estimation system

In the first step, the mission is planned and segmented into several flight and mission phases by the mission objective interpreter. Each of these phases represents a set of tasks. These task models have been developed on the basis of structured interviews during knowledge acquisition experiments with German Army helicopter crews. Based on these sections, the flight phase and mission phase interpreters pursue the progress of the mission, which generates a picture of the current situation. Both, the flight phase and mission phase interpreter have been implemented using state transition networks. This enables the ACU to recognize the task situation in order to deduce the pilots' tasks.

In the next step, the tasks assumed by the ACU, are synchronized with the tasks actually being performed by the human operator. This is done by use of gaze tracking (using faceLAB) as well as observing the manual interactions. Manual interactions which are taken into consideration here include page switching activities on the Control and Display Unit (CDU), further system settings (e.g. landing gear), speech input and output, as well as looking at the magnitude of control stick movements.

During structured interviews with the helicopter crews, every single task has been rated concerning its individual resource demands posed on the pilot. Subsequently, these tasks are transferred into the demand vectors within the pilots' resource model.

Due to restrictions of W/INDEX handling two simultaneous tasks only, the implementation of W/INDEX-model has been modified here by the use of pairwise weighting of all current tasks. Starting with the W/INDEX-conflict matrix suggested by Wickens [14], we modified the conflict-coefficients to our needs.

Based on the estimation of conflicts (caused by actual and possible future assistant system tasks) a rating of preferred interaction modality (either visual-ambient, visual-verbal, auditory-verbal or auditory-ambient) has been implemented. If any information transfer is required, this rating is used by the ACU to choose appropriate OSMs for communicating in a resource adaptive way with the pilot.

5 Preliminary Experimental Testing of Resource Model Prototype

A first engineering test in our flight simulator has been conducted in order to investigate the resource adapted interventions of the assistant system using predictions made by the resource model described in the previous sections.

5.1 Experimental Procedure

In this test, the interventions made by the ACU are of the types "associative assistance" and "alerting assistance" [7]. The purpose of "associative assistance" is to offer its proposals (e.g. "next turn heading") continuously to the pilot such as a commander would do this. In contrast "alerting assistance" assumes to fulfill a kind of artificial cognitive redundancy, which is only intervening in situations, the pilot does not seem to be completely aware of. Here the attention of the pilot is drawn on the most urgent task, e.g. by presenting appropriate hints.

Both, "associative assistance" as well as "alerting assistance" serve to fulfill the goals of cooperation.

Particular emphasis has been put on the selection of the perceptual modality, the ACU chooses in order to interact with the pilot. According to our concept the following hypotheses should be evaluated within the experiment:

- The human operators' remaining resources are dependent on the current task situation.
- Based on the estimated resource consumption, the ACU as assistant system interacts on residual human resources in order to charge these resources equally.

5.2 Mission Configuration

In our scenario, a manned military transport helicopter carries troops from a pickup zone in friendly area to an operation area at foe with two possible drop zones nearby. The commander controls three UAVs, which are flying ahead of the manned helicopter taking over reconnaissance.

Due to this additional UAV-guidance, the pilot is required take over tasks, otherwise typically executed by the commander. These delegated tasks mainly count among mission-management (e.g. radio communication) and routing of manned helicopter during flight.

5.3 Simulation Environment

The mental resource demand predicting prototype has been implemented on a generic two-man side-by-side helicopter simulator, used for research-projects at the Institute of Flight Systems. This simulator consists of six multi-function displays (MFDs), each attached with a touch screen. Depending on the configuration, display formats such as a Primary Flight Display (PFD), a digital map, system status or a radar warning system can be shown. Complex (mission specific) information e.g. communication radio frequency and transponder settings have to be entered into a Control and Display Unit (CDU). For the simulation of the external environment, a three-channel projection system with a viewing angle of 180° was built up. The pilot in command on the left side takes the additional tasks as operator of UAVs.

5.4 Results

In Figure 8, the results of human resource prediction are shown in the context of ACU-interventions. Line 1 to 8 depicts the prediction of resource-consumption, based on the modified VACP. In Line 9, the workload level, estimated by the modified W/INDEX is drawn.

During the flight in friendly area (up to 500sec), the estimated workload was mostly low with only little peaks. The first intervention of the ACU, classified as “associative assistance”, used the visual-verbal channel (VV) to transfer its information to the pilot by generating a text message on display. This modality has been chosen because of allocated demand on the audio-verbal-channel (AV) through a concurrent radio message.

After entering the enemy area (past 600sec) the ACU intervenes by using “alerting assistance” due to the fact, the pilot impends to violate a safety critical aircraft altitude. But by aid of resource adapted interactions, the ACU-induced (not avertable) increase of estimated workload is marginal.

Particularly high resource consumption and workload estimation was observed around 800sec. This correlates with the reconnaissance of enemy jeeps near the desired landing area. In this context, the pilot was working on the rescheduling of the drop zone.

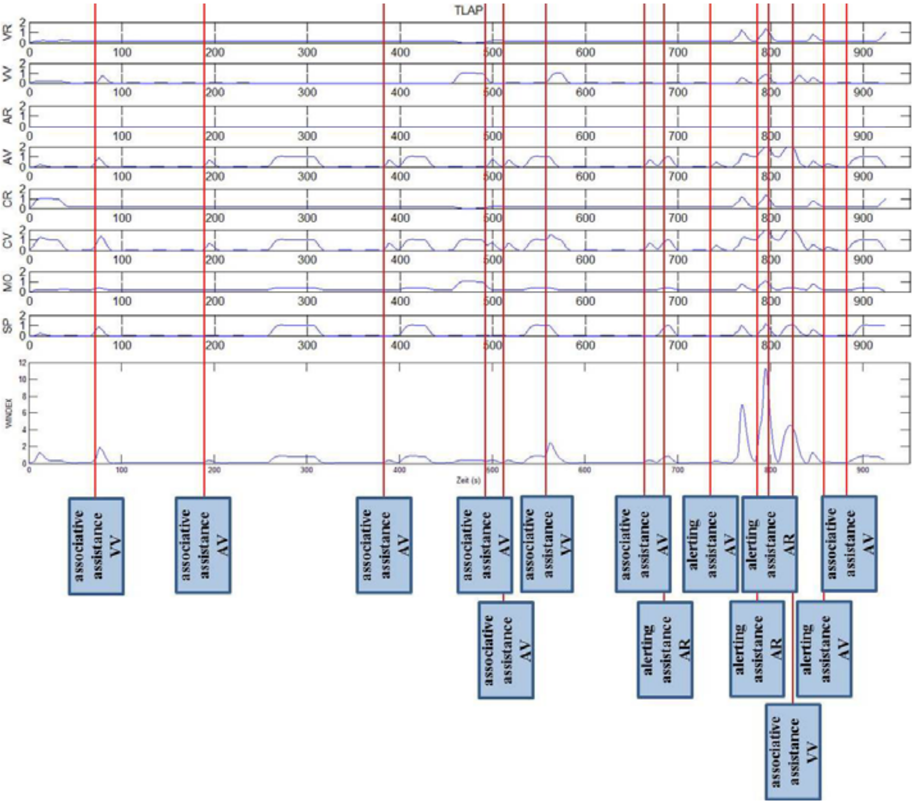


Fig. 8. Experimental results of implemented resource consumption model

6 Conclusions and Perspective

In this article, an approach is presented to enhance cooperative capabilities of future knowledge-based assistant system in the domain of military helicopter missions. For this purpose, we developed a concept for the estimation of pilots' residual capacity in human resources as well as an estimation of his current workload level. By the use of models considering the current resource allocation, an assistant system will be enabled to transfer necessary information on remaining resources of the pilot. This will prevent the pilot from being overtaxed proactively which maximizes the performance of overall system.

The future work will incorporate a validation and verification of the underlying models, especially the task model as well as the resource model, within extensive simulator trails.

References

1. Aldrich, T.B., Szabo, S.M.: A methodology for predicting crew workload in new weapon systems. In: Proceedings of the Human Factors Society 30th Annual Meeting.The Human Factors Society, Santa Monica (1986)
2. Banks, S.B., Lizza, C.S.: Pilot's Associate: a cooperative, knowledge-based system application. IEEE Expert 6(3), 18–29 (1991)
3. Hayashi, M.: Hidden Markov Models for Analysis of Pilot Instrument Scanning and Attention Switching. Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics (2004)
4. Maiwald, F., Benzler, A., Schulte, A.: Berücksichtigung mentaler Operateurzustände bei der Weiterentwicklung wissensbasierter Assistenzsysteme. In: Grandt, M., Bauch, A. (Hrsg.) Innovative Interaktionstechnologien für Mensch-Maschine-Schnittstellen, pp. 303–318. Deutsche Gesellschaft für Luft- und Raumfahrt Lilienthal-Oberth (2010) ISBN 978-3-932182-73-1
5. Miller, C.A., Hannen, M.D.: User Acceptance of an Intelligent User Interface: A Rotorcraft Pilot's Associate Example. In: Maybury, M.T. (ed.) International Conference on Intelligent User Interfaces, Redondo Beach, California, USA, pp. 109–116 (1999)
6. North, R., Riley, V.: W/INDEX: A predictive model of operator workload. In: McMillan, G.R., Beevis, D., Salas, E., Strub, M.H., Sutton, R., Van Breda, L. (eds.) Application of human performance models to system design. Defence Research Series, vol. 2, pp. 81–89. Plenum, New York (1989)
7. Onken, R., Schulte, A.: System-ergonomic Design of Cognitive Automation – Dual-Mode Cognitive Design of Vehicle Guidance and Control Work Systems. In: Studies in Computational Intelligence, vol. 235. Springer, Heidelberg (2010)
8. Prévôt, T., Gerlach, M., Ruckdeschel, W., Wittig, T., Onken, R.: Evaluation of intelligent on-board pilot assistance in in-flight field trials. In: 6th IFAC/IFIP/IFORS/IEA Symposium on analysis, design and evaluation of man–machine systems, Massachusetts Institute of Technology, Cambridge (1995)
9. Ruckdeschel, W., Onken, R.: Modelling of Pilot Behaviour Using Petri Nets. In: Proceedings of the 15th International Conference on Application and Theory of Petri Nets. Springer London, UK (1994)
10. Schulte, A., Donath, D.: Measuring self-adaptive UAV operator's load shedding strategies under high workload. In: HCI International, Orlando (2011)
11. Swezey, R.W., Salas, E.: Guidelines for use in team-training development. In: Swezey, R.W., Salas, E. (eds.), Teams: their training and performance. Ablex publishing corporation, Norwood (1992)
12. Veltman, J.A., Jansen, C.: The role of operator state assessment in adaptive automation. In: TNO (2006)
13. Wickens, C.D., Hollands, J.G.: Engineering Psychology and Human Performance, 3rd edn. Upper Saddle River (2000)
14. Wickens, C.D.: Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science 3(2), 159–177 (2002)

Analysis of Mental Workload during En-route Air Traffic Control Task Execution Based on Eye-Tracking Technique

Caroline Martin[1,2], Julien Cegarra[1], and Philippe Averty[2]

[1] CLLE, Université de Toulouse ; Centre Universitaire,
Place de Verdun-81012 Albi cedex 9, France
julien.cegarra@univ-jfc.fr
[2] DGAC, DTI R&D, 7 Avenue E.Belin, 31400 Toulouse, France
martin@cena.fr, philippe.averty@aviation-civile.gouv.fr

Abstract. This text aims to present a study which deals with mental workload evaluation during task execution. It is focused on the Air Traffic Controllers working situation. In this document, we mainly introduce an experiment which has been conducted in a French En-Route air traffic center with the participation of Air Traffic Controllers. Four principal experiment characteristics are detailed: the experiment procedure, the working situation elaborated for our experimentation, the nature of the task achieved by participants, and the technique chosen to analyze mental workload felt by operators. We finally present the main results from our first data analysis which seem to confirm major observations known in the field of air traffic control, as well as, mental workload study field.

Keywords: Mental workload analysis, Air Traffic Controller, Eye-tracking, eye fixations, pupil diameter.

1 Introduction

A researcher group which comes from the University of Toulouse began, in collaboration with the French civil aviation direction (named DGAC, Direction Générale de l'Aviation Civile), a search project which deals with Air Traffic Controllers (ATCo) mental workload variations during task execution. This project is necessary due to major transformations of air traffic control during the 20th century and especially the further adaptations envisaged during the following years. In fact, assessment of mental workload is a way to evaluate HCI performance [1].

A study has been carried out with the participation of a French En-Route Air Traffic Center. This work's objective is to evaluate "realistic" mental workload felt by ATCo during their work activity. This approach allows analyzing correlations between mental workload variations and air traffic sequence events (like conflict presence). Assessing mental workload can refer to two different approaches: subjective and objective analysis. Subjective mental workload evaluation often consists to fulfill a questionnaire after task execution.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 592–597, 2011.

Objective analysis involves task performance or psychophysiological measures. These latter have for principal advantage to offer the capability for continuous data recording [2] and also for real-time mental workload evaluation. More precisely, we refer to eye movements because it is the human physiological parameter which has the most frequent period of update [3]. We choose to use eye data to analyze ATCo mental workload during task execution in order to obtain detailed evaluation.

2 Method

2.1 Experiment Procedure

The experiment consists in asking ATCo to manage a forty-five minutes air traffic sequence in realistic work conditions. No procedure or action constraint was given to the operator before and during task execution. The experiment objective is to obtain an analysis of mental workload characteristics which transposed realistic aspects of operator's activity.

2.2 Participants

Thirty-seven En-Route air traffic controllers volunteers (9 females and 28 males) were recruited as participants. They all possess the qualification to work on the air traffic sector used for our experimentation. Participants were ranged from 26 to 56 years of age, and reported ATC experience from 0.5 to 26 years.

2.3 Experiment Working Position Characteristics

To carry out this experiment the elaboration of an en-route air traffic simulator has been necessary. Special means have been given to optimize realistic aspect of ATC position elaborated. During experiment, ATCo had at their disposal an experiment working position to realize the requested task. This latter is composed of four main tools:

- Radar screen where the air traffic sequence is projected including real time aircrafts position and air traffic sector boundaries. Operator had the possibility to adjust the radar image parameters (like sector position on the screen);

- Paper strips printed exactly five minutes before each aircraft's sector entry. This small piece of paper contains principal flight information and is used as a quick way to annotate a flight;

- Radio frequency allowing ATCo to communicate with each aircraft pilot of the air traffic sequence. Only one pseudo-pilot played the role of all simulation aircraft pilots. To improve task's realism, pseudo-pilot voice has been modified thanks to a voice distortion system.

- Mouse used by the operator to act on the radar vision parameter settings.

The following figure shows the details of experiment working position.

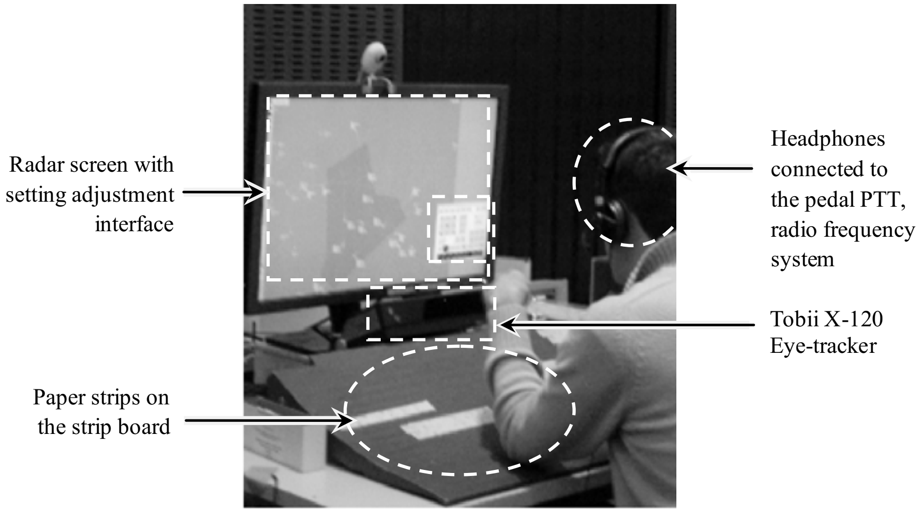

Radar screen with setting adjustment interface
Paper strips on the strip board
Headphones connected to the pedal PTT, radio frequency system
Tobii X-120 Eye-tracker

Fig. 1. The simulated air traffic control position including the eye-tracker

2.4 Air Traffic Task Executed

The air traffic sequence used during the experimentation has been built from real air traffic recorded sequences. We can define the experiment sequence as ecologically valid according to air traffic management point of view because it respects all operational rules applied to control airspace.

The air traffic sequence is composed of fifty-one aircrafts crossing the sector managed by the operator during the experiment. It takes about forty-five minutes.

In the air traffic sequence controlled by participants the number of conflicts varies contrary to number of aircrafts on the radar screen which is relatively constant. Three categories of aircrafts can also be defined: Out of Sector aircrafts (OS), Non-Conflict aircrafts (NC) and Conflict aircrafts (C). Moreover, two situations are distinguished in the air traffic sequence: non-conflict phase (named "Phase 0") and conflict phase (called "Phase 1").

2.5 Mental Workload Evaluation

Mental workload felt by ATCo during task execution has been evaluated with an eye-tracker device (Tobii X-120). Tobii X-120 is a binocular eye-tracker composed of two infrared cameras. We used it with a 60 Hz recording rate. For each experiment a calibration phase has been achieved to optimize quality of recording. Moreover, eye tracker has been placed at 70 centimeters of participant's face because this setting is necessary to ensure data quality. This prerequisite has required to constraint participant's seating position which explains the use of an office chair without wheels.

Eye-tracking technique permits to record in real time the point fixation of gaze, blink periods, saccade trajectories and pupil diameter variations. The eye data analysis here is focused on fixations and pupil diameter. With the help of a dwell-time algorithm we linked the position of eye fixation and the location of each aircraft on the radar screen. This function allows the repartition of the number of fixations according to the several aircrafts categories in order to examine differences between fixations amount by category of aircraft (OS, NC, C).

We also compare pupil diameter according to conflict presence (phase 0 and phase 1). This comparison was focused on two pupil diameter variations [4]: the maximum value obtained during fixations and latency duration (time required to reach the maximum value of pupil diameter). To analyze these data with taking into account inter-individual differences operators concerning their pupil diameter value, we standardized the pupil diameter, which explain the scale observed in the following graphics.

3 Results

Two main results issued of the data analysis can be observed.

3.1 ATCo Attention Allocation

The first one deals with ATCo's attention allocation and is derived from fixations count analysis. In fact, an aircraft hierarchy in terms of attention required is suggested according to the category. Actually, OS aircrafts need significantly less attention (so less fixations count) than other type of aircraft. Moreover, C aircrafts get significantly more fixations count which means more attention by ATCo.

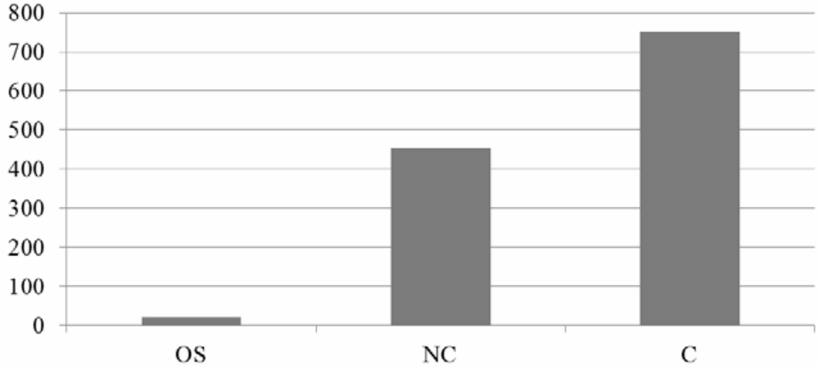


Fig. 2. Changes in fixation count within three several aircraft categories

3.2 Mental Workload Recorded

The second result obtained relates to mental workload. Analysis of maximum pupil diameter shows a significant effect of conflict presence. In fact, pupil diameter is higher when air traffic situation includes conflict (phase 1). This stresses an increase of ATCo mental workload due to conflict presence.

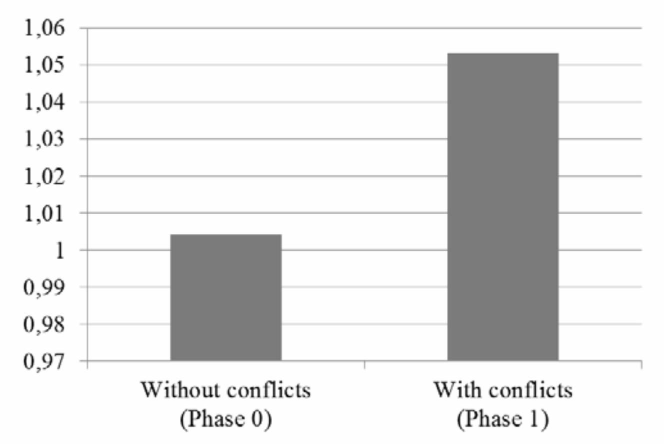


Fig. 3. Changes in maximum pupil diameter according to presence of conflicts (Z score)

Latency analysis allows noting a shorter latency in the presence of conflict, which confirms the effect of conflict presence on mental workload felt by ATCo.

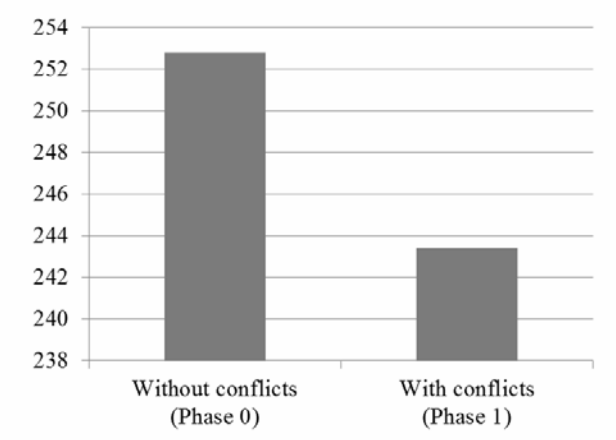


Fig. 4. Changes in latency duration according to presence of conflicts

Currently, complementary data analyses are being carried out. It is focused on the temporal management of workload by ATCo [5]. The objective is to understand the way operators manage their own mental workload level.

4 Discussion

Two limitations of this work have to be highlighted. Firstly, we can consider a limitation in the way mental workload evaluation is achieved because it is focused on only on one physiological source [6]. Indeed, our approach could be completed by another physiological source or by means of a questionnaire/interview with ATCo to obtain ATCo subjective feedback.

Secondly, high frequency of parameter recording wasn't optimally exploited. It could be possible if a temporal data analysis would be carried out. It would allow us to obtain two main signal characteristics: the general trend of pupil diameter value, and the presence of pupil diameter peak, defined as TEPR, Task Evoked Pupillary Response [4]. These findings would allow linking prior mental workload level changes, translating mental workload modification, with events, which happen in the air traffic situation controlled by ATCo.

5 Conclusion

To conclude we can say this study gets hopeful results. Indeed, it allows us to confirm simple hypothesis, which come from ATC field. It particularly highlights the crucial status of conflict in attention and mental workload during ATC task execution.

Results also stress that eye-tracking technique is a powerful approach to study mental workload during a complex activity. Thanks to this experiment we found same results of previous studies carried out with laboratory tasks, especially those showing mental workload growth when task requirements increase. Such results are especially important for guiding design decisions of ATC support systems in the SESAR (Single European Sky ATM Research) project.

References

1. Loft, S., Sanderson, P., Neal, A., Mooij, M.: Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications. Human Factors 49, 376–399 (2007)
2. Kramer, A.F.: Physiological metrics of mental workload: a review of recent progress. In: Damos, D.L. (ed.) Multiple-task Performance, pp. 279–328. Taylor & Francis, Abington (1991)
3. Bridgeman, B.: The Case of Vision Conscious vs unconscious processes. Theory and Psychology 2, 73–88 (1992)
4. Beatty, J.: Task-Evoked Pupillary Response, Processing Load, and The structure of processing resources. Psychological bulletin 91, 276–292 (1982)
5. Averty, P., Collet, C., Dittmar, A., Athènes, S., Vernet-Maury, E.: Mental Workload in Air Traffic Control: An Index Constructed from Field Tests. Aviation, Space, and Environmental Medicine 75(4), 333–341 (2004)
6. Miyake, S., Yamada, S., Shoji, T.Y., Kuge, N., Yamamura, T.: Physiological responses to workload change. A test/retest examination. Applied Ergonomics 40, 987–996 (2009)

An After Action Review Engine for Training in Multiple Areas

Glenn A. Martin, Jason Daly, and Casey Thurston

Institute for Simulation and Training, University of Central Florida,
3100 Technology Parkway, Orlando, FL 32826, USA
{martin,jdaly,cthursto}@ist.ucf.edu

Abstract. The notion of after action review (AAR) is known in the military where it is used to develop a common picture of what happened and why. Recently, the concept has been rediscovered by other domains. Obviously, a review within these domains would be different. This paper addresses development of an AAR engine. By "AAR engine" we mean a system that provides the common functionalities across all AAR systems into a single foundation for training. Regardless of the domain, there are capabilities needed in an AAR system (e.g. recording and playback of scenario data). On the other hand, there are also features specific for each domain. In this paper we first review the infrastructure of our AAR engine. Then advantages of such a system for addressing various AAR systems are reviewed. Additional advanced functions are then presented and reviewed in light of how the engine can easily provide these enhancements.

Keywords: After Action Review, AAR, Simulation, Training, Software Infrastructures.

1 Introduction

The notion of after action review (AAR) is well known in the military domain. Whether squads of dismounted infantry, platoons of tanks or squadrons of aircraft, a post mission review is used to develop a common picture of what happened and why. However, people from other training areas have also been intrigued by what a post simulation review could bring to their training. Obviously, a review within these domains would be different from a military-oriented session.

2 Background

Many AAR systems have been built over the years. In this section we review some of the work by others before covering our own AAR system.

Hixson describes the Corps Battle Simulation After Action Review System [1]. It includes tactical situation displays using information plotted in overlays. A set of

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 598–607, 2011.

military symbology and the capability to make custom symbol sets can also be placed in the overlays. For presentation it includes a spreadsheet and bar and pie charts.

The TACSIM After Action Review User System (TAARUS) is based on field level exercises and specializes in displaying data at levels depending on the audience (e.g. General Officer versus Company Commander) as described in Allen and Smith [2]. The display of maps, graphs, tables and summaries is supported. Key measures of effectiveness include air mission effectiveness, ground sensor performance, trainee intelligence acquisition, air mission tasking timelines and accuracy of intelligence. The display is similar to business graphs (bar and line graphs).

PowerSTRIPES from AcuSoft, Inc. includes a plan view display, maps, charts, graphs, timelines, and data queries [3]. The slides are then viewed during the AAR itself and are available as a take home package.

ViSSA/SA-STAT from ScenPro, Inc. provides a programmable (via a scripting language) capability allowing the operator to define events easily readable from the DIS stream [4]. Events are logged as conditions in the script are satisfied and these events are available for use in the AAR session. It is one of the most configurable systems although still has a military focus.

AutoCAS from Aptima, Inc. focuses on the voice communication within a training scenario [5]. All the voice communication is recorded and analyzed including a voice recognition transcription capability. A timeline showing when each user's utterances were spoken is included. In addition, a hierarchical command structure display showing when a leader is speaking "up" or "down" the command chain is also provided. The communication analysis that AutoCAS supports fills a good hole in that much information is encoded in the utterances of each team members.

Jensen et al discusses the CACCTUS AIRS system [6]. This AAR system is designed upon causal explanation analysis. Three types of errors are supported: procedural errors (which are skill based and easily defined), cognitive errors (which are knowledge based and show the level of understanding of concepts), and unintended errors (which are based on so called "human errors" – just simply mistakes that humans make from time to time). The error data are used to provide events for the AAR review facilitator to discuss. The AIRS system has a nice theory in its development but is currently military focused.

Most of these AAR systems, however, are focused to their domain and have only minor configurability. Extending them to a new domain can be difficult.

2.1 DIVAARS

Knerr et al have previously built the Dismounted Infantry Virtual After Action Review System (DIVAARS) as a system for capturing and replaying segments of an exercise with a focus on dismounted infantry [7]. DIVAARS includes full DVD-like capabilities along with visual augmentations (bullet lines, movement trackers, etc.) to aid understanding. Figure 1 shows an example of a scenario with movement tracks and field-of-view indicators turned on.

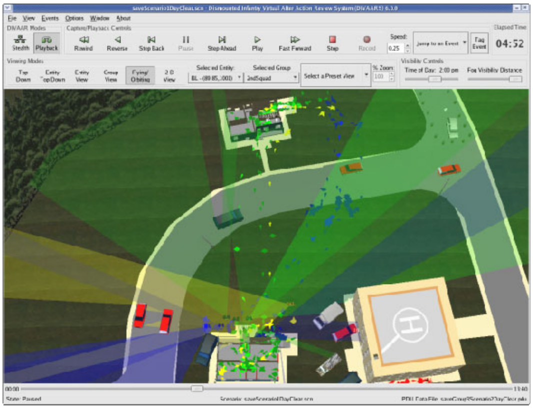


Fig. 1. DIVAARS

We have recently extended the capabilities of DIVAARS to handle Fire Support Teams (FiST). While still a military-based system, FiST scenarios vary quite considerably from the urban combat focus of DIVAARS. Additional visual augmentations are necessary (e.g. artillery arcs and missile paths). Figure 2 shows such a scenario in a specialized DIVAARS, which we named the Joint Terminal Attack Controller After Action Review System (JTACAARS).

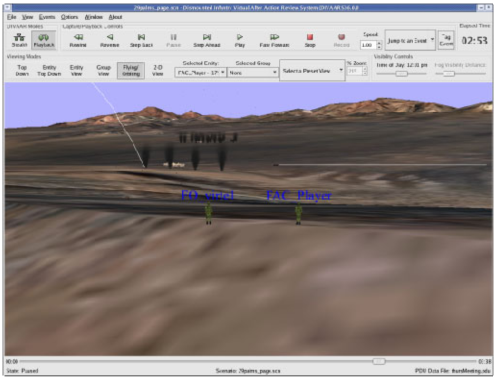


Fig. 2. JTACAARS

2.2 Other AAR Systems

In addition, the general concept of the AAR has been "rediscovered" by domains other than the military. Fidopiastis et al have recently been using modeling and simulation as a new method to increase cognitive ability in brain injury patients [8]. As a part of this they desire the ability to sit down and review each patient's performance whether in mock scenarios (such as a mock kitchen as a part of a rehabilitation clinic) or in virtual/augmented environments.

In addition, Dieker et al have used simulation in reviewing how prospective teachers deal with classroom scenarios such as handling an unruly child [9]. Bringing AAR to this training, an observer would sit down with the teacher during the review process to discuss what they did, why, and what could be done differently. In short, the concept of AAR has wide use potential in many domains beyond the military. Indeed, AAR is where the training occurs. A simulation can provide a method to practice, but the review process is the key to understanding.

3 An After Action Review Engine

In order to address performing after action reviews across many domains, we have built the System of Object-based Components for Review and Assessment of Training Environment Scenarios, or SOCRATES. We did not build it as a new general-purpose AAR system, but rather as an AAR engine. By "AAR engine" we mean a system that provides the common functionalities across all potential AAR systems into a single foundation for all training environment scenarios (much like a game engine provides the common needs of games).

Whether military or another domain, there are many capabilities needed in an AAR system. For example, recording and playback of scenario data with full ability to pause, rewind and jump to specific events. In addition, support for a graphical user interface and potentially other interfaces (such as an electronic whiteboard) could be included.

On the other hand, there are also some features that may be specific for each domain. A military scenario may have tabular data such as who killed whom. A cognitive rehabilitation scenario may need to track objects to identify a patient that easily loses attention. A teaching scenario may need to automatically mark events caused by an unruly student.

3.1 Structure

To address these issues the AAR engine has a fundamental architecture of common functionalities coupled with a plug-in architecture for specific domain features. In addition, the plug-in architecture allows users to write their own new capabilities to include in their AAR. Furthermore, they could even write a set of plug-ins to create an AAR system for their own, new domain.

In order to allow the core functions and the plug-in modules to communicate, an event system was built to connect all of the components together. Events can be both issued and received by each component. Some events are well-known that all components may need to process (such as a "Play" event) while others may be specific to a subset of modules and all other components will ignore them. Figure 3 shows a diagram of the architecture with example modules in place for what would exist for our dismounted infantry AAR system (DIVAARS) implemented on top of SOCRATES.

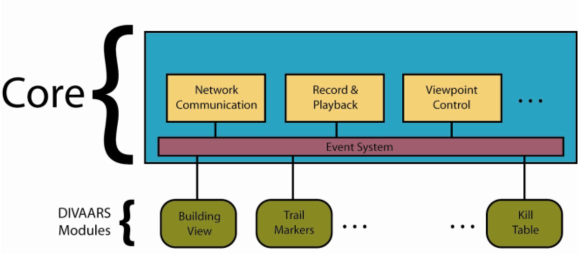


Fig. 3. DIVAARS on top of SOCRATES

As suggested, such an architecture allows other modules to be loaded to essentially create a new after action review system. One such new area is within cognitive rehabilitation. We are in the process of developing the Application for Cognitive Observation and Remediation Needs (ACORN) that provides the ability to record and review a session with a patient in a cognitive rehabilitation setting. The architecture for ACORN might look like Figure 4 (note the similarities and differences to Figure 3).

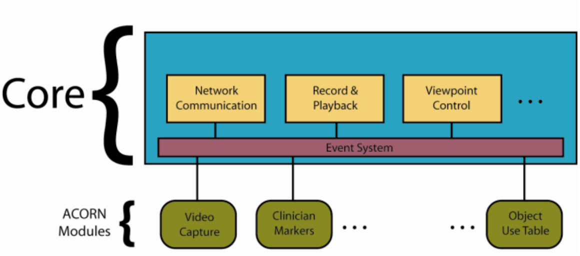


Fig. 4. ACORN on top of SOCRATES

A system for reviewing educator training would look similarly.

In addition to providing flexibility, the plug-in architecture also provides the capability not to load a feature. If a particular review system is being used on a less capable machine (such as a wearable computer), some features could be disabled. When improved machines are available, these features can simply be re-enabled by loading the plug-in module once again.

3.2 "Thoughts"

In order for the plug-ins to communicate with each other, an event message passing scheme is used. We refer to each message as a "thought" to avoid confusion with scenario events that may occur during the training scenario.

Each plug-in can register for thought types as well as issue new thought instances. As of now, registration is only supported by type (for example, all "Play" thoughts); however, registration by other filters could be an element of future work. During each system loop each plug-in queries for pending thoughts and processes each accordingly.

In order to provide the capability of different plug-ins running at different speeds, the AAR engine presented here supports multiple threads (definable in a configuration file). Each plug-in is added to one thread and each thread runs at a speed given in the configuration file. This allows a plug-in such a renderer or network packet processor to run faster than the user interface, for example.

To avoid conflicts at start-up, each plug-in is initialized in turn and are disallowed from issuing thoughts until all plug-ins are initialized. Without this restriction, some race conditions of scenario data and visualization data are possible. Therefore, a rolling 2-component set of barriers is used where each thread of plug-ins releases the next thread for initialization.

3.3 Plug-ins

In this section, some of the more important plug-ins are discussed. While more than fifty plug-ins exist in our system to date, covering every single one is beyond the scope of this paper. However, we do cover some fundamental and basic plug-ins as well as two complex ones.

Fundamental Plug-ins. The GUI (Graphical User Interface) plug-in handles all aspects of the user interface. It contains core elements that are common to all AAR systems such as the DVD-like controls and the tagging of exercise events. In addition, it provides a capability where other plug-ins can register GUI components of their own. This allows the other plug-ins to provide controls to the user. For example, display of fog and time of day can be toggled on and off.

The Network plug-in handles reading the simulation data from the network and issues appropriate thoughts based on the exercise. In addition, there is a corresponding Commo plug-in that does the same for voice communication. Having the two separate plug-ins allows each to run at different speeds. For example, the Commo plug-in runs faster in order to get all the voice data issued in a timely fashion (voice is particularly susceptible to added latency).

The Sim Logger plug-in saves all simulation and voice data when in a recording mode. Similarly, the Log Playback plug-in plays back a saved file during playback mode. We chose the eXtensible Mark-up Language (XML) to store the saved data. XML is useful for hand editing (if necessary) and provides a flexible format for others to read. We have worked with partners that perform analysis of the exercise data and having an XML-based format makes our partners' work easier.

Similar to the Sim Logger/Log Playback plug-ins, the Run Info plug-in stores all AAR-specific data during a run. This includes events during the exercise that the user has tagged (both a time and a textual description is saved). Run Info also stores information for plug-ins that register with it. For example, a trail marker plug-in can store the points making up an entity's trail with the Run Info plug-in so that it does not need to be re-computed each time.

Other Basic Plug-ins. While not as necessary as the before-mentioned plug-ins, the Telestrator plug-in is an example of the structure of the system coming together to provide a function. It is designed for use by the AAR operator, allowing him or her to highlight specific objects or elements in the scene. Much like a sportscaster can augment his commentary by drawing on the screen, the Telestrator plug-in allows the

AAR operator to draw on the visual scene in order to direct the audience's attention to specific visual elements or to pictorially describe the events that occurred or are about to occur. The Telestrator supports several basic shapes as well as freehand drawing. The operator can pick from several colors and line thicknesses. The Telestrator was originally designed to support distributed AAR sessions (described below), but it has proven useful in standalone sessions as well.

Extension-based Plug-ins. Certain plug-ins by nature are themselves modular, and can benefit from a modular architecture, much as the overall AAR system is composed of plug-ins. We refer to these modules of a plug-in as "extensions." The extension system is provided by the SOCRATES core, providing a specific framework, and allowing extension-based plug-ins to reuse the code that makes extensions possible. So far, we have built two plug-ins that utilize this capability. These are the Virtual Renderer and Metrics plug-ins.

The Virtual Renderer plug-in creates a three-dimensional visual representation of the training exercise, allowing the user or users to visually see the events that are taking place (or previously took place). The plug-in is relatively complex and responds to a wide variety of thoughts. Aside from the standard control thoughts (play, record, pause, etc.) it listens for thoughts that transmit the state of "elements" (entities or participants in the exercise), as well as other more specific thoughts that deal with simulation events. Certain thoughts are handled by the plug-in itself, such as the creating and positioning of windows or viewports on the GUI, and the loading of the static 3D environment. The response to other thoughts can depend on the type of exercise being conducted. This is where the modularity of the Virtual Renderer pays off.

The Virtual Renderer is configured much like a standalone image generator, providing basic visualization capabilities that can be accessed and utilized by other plug-ins. The visualization of entities is handled by the Element Manager Extension, which is the most complex extension. This extension is highly configurable, allowing the specific application to tailor how its exercise elements are visualized. The remaining extensions are designed to handle specific visual augmentations. One extension can draw text in the scene, which can be used to label entities or specific locations. Another can draw connected line segments, which can be used to show the path that an entity or object takes through the scene. Other extensions provide for drawing fading line segments (which can be used to show weapons fire), placing two dimensional images in the 3D scene (which provides alternative visualizations of elements), and adjusting the scene visibility parameters such as lighting and fog. A few extensions are specific, such as the extension that supports the telestrator plug-in.

The Virtual Renderer does make the assumption that the exercise in question can and should be visualized using 3D computer generated imagery. This may not be appropriate for all kinds of exercises or after-action reviews. Another kind of renderer could be a video renderer that simply records and replays one or more streams of video from cameras that are positioned to record the essential aspects of the exercise. These streams could then be replayed in sync with the rest of the simulation data during the AAR session. Like the Virtual Renderer, this video renderer could conceivably add its own visual augmentations to the video playback (highlighting certain elements in the scene, or tracing the path of some object). In fact, the video renderer could register and react for the same thoughts but simply render them differently (e.g. movement tracks could be overlaid onto the video, etc.).

Another plug-in that benefits from modular design is the Metrics plug-in. This plug-in is designed to analyze the events that take place during the exercise (including the movements and actions taken by the exercise participants), and generate specific measurements from these data. In effect, the metrics plug-in provides a different kind of view of the exercise. Where the Virtual Renderer provides a real-time visualization, the Metrics plug-in provides a detailed timeline of the exercise's significant events, and a summary of how well the participants carried out their tasks. It can identify failures in procedure or execution, and even point out the specific wrong action taken that led to subsequent failures.

This kind of detail requires very specific knowledge of the exercise. The Metrics plug-in was originally written to support experiments dealing with fire support team training, and required over 150 specific metrics to be calculated. It collects and manages simulation data in its core, but the metrics themselves are calculated in extensions. This design keeps the source data in one place (avoiding duplication), but allows it to be used for many different calculations. The modular design also allows the plug-in to be used for purposes other than the original experiment. The application can specify which metrics should be active in its configuration, and new metrics can be created simply by writing additional extensions.

3.4 Special Capabilities

The plug-in system allows some special capabilities to exist in a relatively straight-forward manner. Here, we cover two in particular: the remote interface for additional AAR data, and a distributed AAR capability.

Remote Interface. We work with many partners in the area of after action review. These partners analyze the same simulation data to perform more in-depth review of performance. For example, some may focus on situation awareness or voice communication analysis. With these tools they may find key events that should be reviewed or they may create new data for display during the AAR.

DIVAARS provides a "remote interface" where other AAR analysis applications can transmit new tagged events or tabular data for use during the AAR session. The AAR engine implements this by a RemoteInterface plug-in that accepts network connections and receives commands that represent the new events and new tables. These are then simply forwarded by the ThoughtSystem for other plug-ins to handle normally through the use of a TaggedEventThought and a StatTableThought, respectively.

Distributed AAR. We have been basing most of our current research in the area of game-based training. Specifically, we are using existing game-based training systems in a distributed exercise. To perform AARs, we have built a client application known as the Distributed AAR Remote Toolkit (DART). DART is built on top of SOCRATES but runs as a client participating in a distributed AAR session.

DIVAARS and DART work together in a distributed AAR. Each DART instance connects to the master DIVAARS station with the AAR Facilitator using DIVAARS and each trainee using DART. The users can then use an AARIntercom plug-in that provides voice communication during the AAR session and the Telestrator plug-in to provide drawing capabilities.

In order to provide the connectivity of the DIVAARS and multiple DART instances, we created two plug-ins. The DistributedServer plug-in registers for thoughts that the clients will need and handles TCP client connections. As each registered thought is received, this plug-in gets it in XML form and transmits it across the TCP link. Similarly, it receives XML from the client connection, creates a thought based upon the XML and issues it (then all other plug-ins registered for this new thought will receive it as normal).

Similarly, a DistributedClient plug-in was also created that the DART application uses. It makes a TCP connection to the master DIVAARS station (i.e., to the DistributedServer plug-in on that station), and, similarly, transmits XML for registered thoughts and processes incoming XML into issued thoughts.

Since thoughts know how to output themselves as XML and to create themselves based upon XML, creating a distributed AAR capability was very straight-forward and shows the flexibility of the plug-in architecture. As new thoughts are created for new functionality, the DistributedServer and DistributedClient plug-ins can be modified to register for the corresponding thoughts to pass that capability between each other. Similarly, DART can be modified to handle a distributed AAR capability for any domain, not just a distributed military domain.

4 Conclusions and Future Work

After Action Review is well known within the military and is seeing expanding interest in the civilian world. Development of a system to address the varying needs of each domain provides a key platform to apply the process of after action review to multiple domains. The AAR engine presented here provides sufficient capabilities, yet flexibility, to address such a solution. In addition, an AAR engine can provide specialized capabilities such as distributed AAR and a remote connection interface in an easy fashion.

Acknowledgements. The authors would like to acknowledge the support of the Army Research Institute (ARI) and the Office of Naval Research (ONR) who helped partially fund this work. In addition, we would like to thank all the developers who helped implement many of the plug-ins in SOCRATES.

References

1. Hixson, J.A.: Using After Action Review Systems for Exercise Planning and Control. In: Winter Simulation Conference (1996)
2. Allen, G., Smith, R.: After Action Review in Military Training Simulations. In: Winter Simulation Conference (1994)
3. PowerSTRIPES, http://www.acusoft.com/products/powerstripes/
4. ViSSA+, http://www.scenpro.com/p_vissa.html
5. AutoCAS Audio Communication Analysis Tool, http://www.aptima.com/

6. Jensen, R., Chen, D.Y., Nolan, M.: Automatic Causal Explanation Analysis for Combined Arms Training AAR. In: Interservice/Industry Training, Simulation and Education Conference, Orlando, FL (2005)
7. Knerr, B.W., Lampton, D.R., Martin, G.A., Washburn, D.A., Cope, D.: Developing an After Action Review system for Virtual Dismounted Infantry Simulations. In: Interservice/Industry Training, Simulation and Education Conference, Orlando, FL (2002)
8. Fidopiastis, C.M., Stapleton, C.B., Whiteside, J.D., Hughes, C.E., Fiore, S.M., Martin, G.A., Rolland, J.P., Smith, E.M.: Human Experience Modeler: Context Driven Cognitive Retraining and Narrative Threads. 4th International Workshop on Virtual Rehabilitation (2005)
9. Dieker, L., Hynes, M., Stapleton, C.B., Hughes, C.E.: Virtual Classrooms: STAR Simulator. In: New Learning Technologies, Orlando, FL (2007)

Mixed-Initiative Multi-UAV Mission Planning by Merging Human and Machine Cognitive Skills

Ruben Strenzke and Axel Schulte

Universität der Bundeswehr München (UBM), Department of Aerospace Engineering
Institute of Flight Systems (LRT-13), 85577 Neubiberg, Germany
{ruben.strenzke,axel.schulte}@unibw.de

Abstract. The Universität der Bundeswehr München is conducting research in the field of Manned-Unmanned Teaming (MUM-T). In the MUM-T scenario there is a human multi-UAV (Uninhabited Aerial Vehicle) operator who is responsible for the online air mission planning and re-planning. This operator shall be supported in his work by an assisting automation in order to maximize system performance. We therefore examine multiple scientific approaches to human-automation integration and present our established Cognitive and Cooperative Automation approach as well as a novel Cognitive Skill Merging approach. The latter is based upon bringing together human and machine cognitive skills in order to cooperatively reason about and work upon the common overall mission planning task without decomposing it in advance. The combination of these approaches results in the proposal of applying mixed-initiative planning to address the above-mentioned problem. The concept of the MUM-T Mission Planner is presented and future experiments are outlined.

Keywords: multi-UAV, mixed-initiative, mission planning, assistant system, manned-unmanned teaming, human-machine interaction, cognitive automation, cooperative automation, artificial intelligence, human-automation integration.

1 Introduction

Uninhabited Aerial Vehicles (UAVs) in use today are typically guided by a crew of at least two human operators. There are approaches to invert the operator-to-vehicle ratio in the future [1] [2]. For this reason the Universität der Bundeswehr München (UBM) is conducting research on semi-autonomous UAVs using the Cognitive and Cooperative automation approach [3] in order to cope with the high work demands.

This article first gives an overview of the problem posed by the MUM-T scenario (chapter 2). In chapter 3 different existing human-automation integration philosophies are evaluated with respect to our MUM-T application and the Cognitive Skill Merging MABA-MABA (Men Are Better At – Machines Are Better At) scheme will be introduced. As a result, we propose the assistance of the human operator via mixed-initiative planning, which is based upon human-automation integration. In chapter 4 the article explains the MUM-T Mission Planner (MMP) concept and chapter 5 finally gives an outline of the future human-in-the-loop experiments.

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 608–617, 2011.

2 Multi-UAV Mission Planning Application

In this section we first describe the MUM-T scenario and the UAV operator workplace. Then, we derive the need of an assisting automation for the human operator.

2.1 The Manned-Unmanned Teaming Scenario

The MUM-T mission domain refers to a time-constrained multi-aircraft mission including the aspects of transportation, reconnaissance and high-level path planning [2] [4]. The UBM is running a research helicopter simulator with a commander and a pilot workplace to demonstrate and test such missions. This simulator has been used for MUM-T experiments already in which the helicopter commander had the responsibility to guide multiple UAVs (e.g. [2] [4]). For our mixed-initiative planning approach we examine a single operator multi-UAV scenario in which the operator has a simple moving map-based planning interface allowing him/her to create mission plans (cf. Fig. 1). In our default mission there need to be given up to 12 individual orders to each UAV, which in the case of a three UAV mission results in approximately 36 UAV orders in total. These orders are specified by a type (e.g. route reconnaissance, area reconnaissance, object surveillance), a target location and the designated UAV. Each order is inserted into a sequential agenda. Afterwards, it is automatically connected with the preceding action and the successive in case there is one. It is not possible to specify anything like an execution or arrival time for the action. Hence, the UAV operator has not only the responsibility to generate more than 30 orders for the three UAVs at the beginning of the mission but he also has to check if the mission time constraints (e.g. ingress and egress corridor time windows) will be met with his/her plan. The same applies to all re-planning situations that arise during the mission. In these cases there is also a time pressure for the operator. More details to the task-based guidance concept used in the MUM-T application can be found in [4].

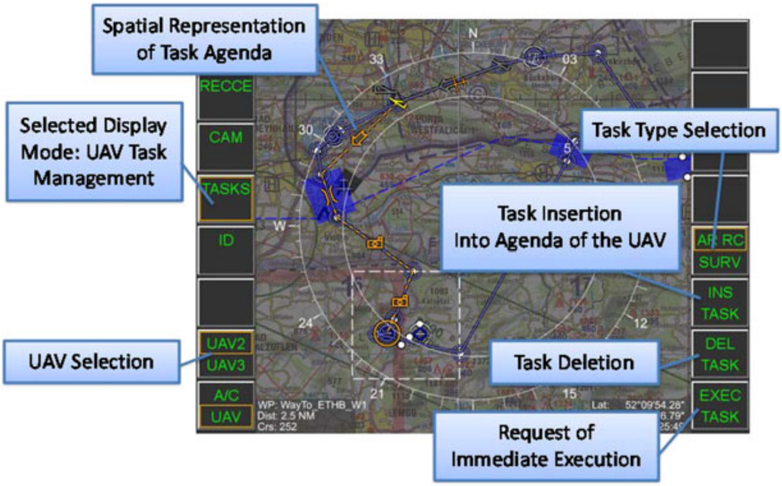


Fig. 1. Task-based UAV guidance graphical user interface

2.2 The Need for Operator Assistance

The problem that we are addressing in the MUM-T application is to maximize overall mission performance. An important aspect is the optimal support of the manned helicopter by the reconnaissance UAVs flying ahead. To grant this support the human UAV operator has to task the UAVs correctly. This means, he has to generate an optimal mission plan at the beginning of the mission (then working under low time pressure) and also has to maintain it throughout the mission (possibly requiring time critical decision-making). Therefore, the operator has to react to unforeseen events quickly and thoroughly by re-planning the mission according to the new requirements (e.g. the primary landing site is threatened; reschedule all UAV-orders to the alternate landing site). In case the operator makes a mistake while entering orders into the system it is likely that he will notice the unwanted result quickly due to the feedback on the graphical display (see Fig. 1). This leads to the necessity to edit or deleted UAV actions, further increasing the time pressure.

Therefore, the first idea could be to implement an automated planning system that solves the problems mentioned above and thereby reduces the operator's workload. But there are certain conceptual and implementation-related problems concerning the application of a fully automated mission planner. First of all in such a complex planning problem it is very difficult even for state-of-the-art planning engines to find an optimal or at least sufficiently cost-efficient plan in a reasonable response time. But even if we had algorithms which work fine in all possible situations of our real-time simulation or assuming we had unlimited processing power, there still would be the need for the operator to check and understand the plan(s) generated by the automation. This is because he/she has the highest authority concerning mission planning due to his/her responsibility for the mission. Therefore, he/she cannot risk not being deeply involved in the decision, i.e. not knowing the mission plan and not knowing its implications. Price [5] also speaks of cognition starving in such cases. But this is not the only reason to integrate the human into the mission planning process, although it should be seen as a sufficient one. Another important aspect is that the human operator contributes intuitive probabilistic reasoning and solution search heuristics, which are very difficult (if not impossible) to either extract or operationalize or implement into an artificial intelligence system. An example for this is the operator's plan to use a certain number of UAVs for the mission. Although a lower number of UAVs would be sufficient to fulfill the mission goals he decides to take a reserve with him for the case that unforeseen problems arise during mission execution (e.g. loss of a UAV, emergence of enemy forces, follow-up orders from mission command).

3 Human-Automation Integration Approaches

The problem of human-automation integration classically deals with the question which functions to automate and which to leave to the human operator. In this section we first give an overview of different classical and modern approaches. After that, we describe our Dual-Mode Cognitive Automation approach and a *Cognitive Skill Merging MABA-MABA* (*CSM3*) scheme as well as the mixed-initiative paradigm.

3.1 Function Allocation Strategies

S*tatic function allocation* strategies distribute the tasks of the human and the machine during design time. There are different approaches that either maximize automation as such or try to minimize the economic costs. More human-focused strategies are to establish the lowest possible grade of automation or the static allocation of tasks in correspondence with the abilities that the human and the machine player bring into the game. The latter is based on Fitts' List and is also called MABA-MABA (Men Are Better At – Machines Are Better At) or compensatory principle due to compensation of human as well machine weaknesses. If we applied a static function allocation strategy to the MUM-T mission planning task as a whole this would result in either fully automated planning or fully manual planning. Both solutions have previously been excluded (see chapter 2). Another possibility would be to decompose the mission planning task into subtasks and allocate these to the human and the machine. We will show in section 3.3 why this is a problematic approach.

D*ynamic function allocation* or adaptable automation is a reaction to the fact that today tasks are dynamic. Therefore their allocation should also be handled dynamically [6]. Systems following this approach are able to dynamically distribute tasks among human and machine during run-time. If we applied a dynamic function allocation strategy to the MUM-T mission planning task as a whole this would result in fully automated planning and fully manual planning alternating or again end in a (dynamic) task decomposition. Therefore the same critique applies to this approach as to the static function allocation.

3.2 Cognitive and Cooperative Automation

A human factors framework supporting the examination of advanced automation is the concept of the work system [3]. The work system consists of two major elements, the Operating Force (OF) which is the high-end decision component that pursues the overall work objective and the Operation Supporting Means (OSM) that are applied by the OF to accomplish the work objective. Both are combined in order to achieve a certain work result on the basis of the given work objective while being constrained by environmental conditions (e.g. information and resources). The OSMs have no knowledge of the overall work objective and perform the assigned subtasks that the OF derived from the work objective. This relationship can be described by the *supervisory control* paradigm [7].

Cognitive Automation is realized by so-called Artificial Cognitive Units (ACUs) which simply spoken are artificial knowledge-based agents. According to Onken and Schulte [3] there are two ways to introduce Cognitive Automation into the work system, either as part of the OSMs or the OF. An ACU being part of the OSMs is called *Supporting ACU* (SCU). An ACU that is part of the operating force is called *Operating ACU* (OCU) or cognitive assistant system. In contrast to an SCU it knows and understands the work objective. The operator and the OCU form a team (*Cooperative Automation*) and can both derive necessary tasks from the common work objective as well as delegate tasks to the available OSMs, both driven by own initiatives. Considering both types of implementations of Cognitive Automation, SCUs and OCUs, is what we call *Dual-Mode Cognitive Automation* (cf. Fig. 2).

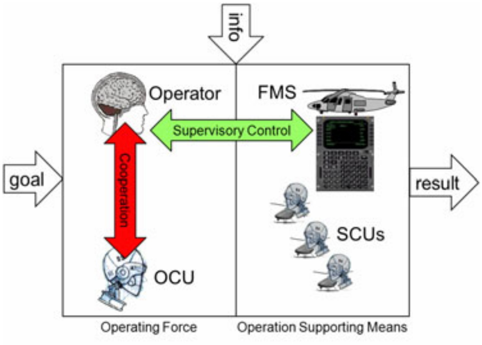
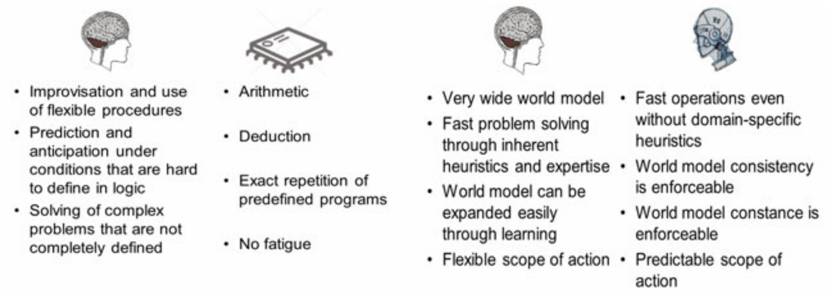


Fig. 2. The MUM-T work system and the Dual-Mode Cognitive Automation approach

3.3 Cognitive Skill Merging MABA-MABA

Humans and machines have different cognitive skill sets and even with artificial intelligence technologies steadily advancing this will still be true in the future. Therefore, MABA-MABA-like schemes can become interesting again as soon as the viewpoint is changed towards true human-machine cooperation instead of task distribution among human and machine. Task allocation strategies cannot cover machine contributions like giving the human a hint concerning a specific task. Such an example can be better circumscribed as *sharing of control* [7] with both human and machine working and reasoning upon the same task and cooperating. In contrast, task allocation can be seen as *trading of control* [7]. The sharing aspect includes extending the humans capabilities (making him/her perform better than he/she could without automation support) and relieving the human of some of the work. Relieving does not have to be realized in form of task re-allocation; instead it is also possible for the machine to only comment on user actions and thereby avoid future human errors and associated clean-up tasks. Such kind of machine assistance can be circumscribed as a *virtual teammate looking over the operator's shoulder*. Although the sharing concept exists for a long time now it seems to be picked up very rarely. One might argue that the sharing of control automatically arises as soon as a task is decomposed into subtasks and these are then distributed among human and machine. But it is important to note that the decomposition of a task, which generates smaller sub-problems for either the human or the machine to solve, brings complications with it as soon as we want to apply the Cooperative Automation approach. As stated before, the knowledge about the work objective is crucial for true Cooperative Automation. Reasoning upon the complete work objective and all available parameters is the strength of an OCU. Therefore, there has to be at least one processing layer in the machine that works upon a non-decomposed task. In our MUM-T example this task is in the best case the overall mission planning task. On the other hand, dynamic subtask allocation may cause out-of-the-loop effects for the human operator due to his/her concentration on disjointed subtasks. Instead human and machine should cooperate and communicate with each other while they reason and work upon the same problem or task. This is very similar to the conclusion of Dekker and Woods [8] that the main question for successful human-automation integration is not about which tasks are automated at which level but how human and machine cooperate.

Hoyos [9] released a revised MABA-MABA list and we selected the abilities that are relevant for the planning application that is described in this article, which leads to the list shown in Fig. 3 (left). We would like to comment on the strengths of the machine in order to derive our Cognitive Skill Merging MABA-MABA (CSM3) scheme in the following. The first one is *arithmetic*, which is an undisputed strength. What it is good for in our application field is the calculation of exact costs and times during either plan generation or plan evaluation. The second strength is fast and infallible *deduction* which is also very advantageous in machine reasoning and automated planning. But this deduction can only take place inside a predefined and more or less limited model of the world. There are numerous reasons for this limitation but the most important and simple one is that someone (e.g. system developer) has to specify the world model and there is only somehow limited time for this modeling process. The third strength of the machine is the *exact repetition of predefined programs*. This ability also has an implication on the machine way of thinking (use of a world model), for it allows consistent reasoning upon the world model. Finally, the machine knows *no fatigue*, which implies that in theory no errors will occur while it is reasoning upon its world model, independent of for how long the machine is doing this.



- Improvisation and use of flexible procedures
- Prediction and anticipation under conditions that are hard to define in logic
- Solving of complex problems that are not completely defined

- Arithmetic
- Deduction
- Exact repetition of predefined programs
- No fatigue

- Very wide world model
- Fast problem solving through inherent heuristics and expertise
- World model can be expanded easily through learning
- Flexible scope of action

- Fast operations even without domain-specific heuristics
- World model consistency is enforceable
- World model constance is enforceable
- Predictable scope of action

Fig. 3. Selected abilities of Hoyos' MABA-MABA list (left) and the CSM3 scheme (right)

As described above the machine's world model plays an important role in its performance and it therefore plays an important role for understanding machine strengths when speaking of artificial intelligence systems. And we have shown above that when using Cognitive Automation, Cooperative Automation and mixed-initiative planning systems there is always artificial intelligence involved. In Fig. 3 (right) we therefore present what characterizes each the human and the machine individually when they both reason and work cooperatively on the same problem or task. We call this scheme Cognitive Skill Merging MABA-MABA (CSM3) because human and machine therein merge their diverse cognitive skills, human cognition on the one side and artificial cognition on the other. Due to cognitive abilities the human has a very rich world model which grants him the improvisation capability and flexibility mentioned in the MABA-MABA scheme. In theory, the richer the world model is the more possible solutions a human can find to a problem. This does not automatically count for a machine as well. Inferring over a larger knowledge base can take longer and having more solution possibilities for a problem can lead to a more demanding (longer lasting) search for the best solution (in terms of costs). The human does not have these two problems because among other things he follows problem solving heuristics that are either inherent or learned (like expert knowledge, experience and expert intuition). This also means that the human's model of the world can be modified over time. Most of the time we mean an expansion of the world model when we speak of learning but also modifications and forgetting are possible. Machine learning on a symbolical level is a very complicated issue. We therefore see the strength of the machine rather in being able to maintain a constant world model if this is intended by the developer. In contrast to this, the human cannot switch off his learning function by any means. This makes machine behavior predictive and human behavior flexible. But it is important to note that the machine behavior in terms of a response to a given problem does not have to be deterministic. It only means that the alternatives in the sense of atomic actions or parameters are predictable, but the complete plan or answer that a machine generates can be a complex structure out of these atomic building bricks. Another advantage of the machine is that a consistency of the world model can be enforceable. This is of course easier for a smaller world model. In contrast, the human's model can be inconsistent, i.e. it can contain contradictions and can also be inconsistent over time. Finally, it is a strength of the machine to search problem spaces with brute force, random search or domain-independent heuristics and therefore find solutions in a domain for which it has no expert knowledge. As an example, for such a problem a human expert could be faster in solving it (due to his expert heuristics) and a human non-expert could be slower.

By using the example of the MUM-T work system (cf. Fig. 2) and applying a static (skill dividing) MABA-MABA task allocation approach to the UAV (OSM/SCU) guidance and applying the (skill merging) CSM3 approach to the team consisting of the human operator and the assistant system (OCU) we obtain the configuration shown in Fig. 4.

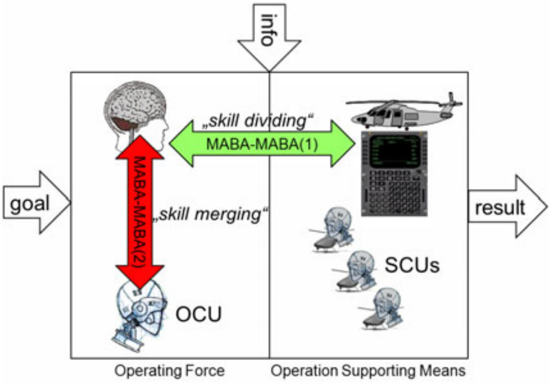


Fig. 4. Static function allocation and CSM3 applied to the MUM-T work system

3.4 The Mixed-Initiative Paradigm

Tecuci, Boicu and Cox [10] define the mixed-initiative paradigm as human and automation cooperating to achieve a common goal. This is similar to the Cognitive Automation approach described above. Furthermore, the mixed-initiative systems are designed either to accomplish goals that are unachievable by the human or the machine on its own or to increase system effectiveness. It is important to note that Tecuci, Boicu and Cox do not speak of task allocation but instead of an interleaving of contributions by the human and the machine which have different knowledge and different skills. These contributions are dynamic as far as their content and point of time are concerned. Of course not all systems that have mixed-initiative functionalities are mixed-initiative planning systems. But many of them are associated with planning and/or scheduling tasks. Some of these systems are also in real use [11]. Comparing the human and machine responsibilities and the methods of human-machine interaction in all these systems shows a heterogeneous field. A very good example that is close to our view of what mixed-initiative planning is (and how it should be, in accordance with the CSM3 scheme and sharing of control) is the Rochester Interactive Planning System (TRIPS) [12].

4 Manned-Unmanned Teaming Mission Planner Concept

The system that is designed to support the human operator in the mission planning process via mixed-initiative is called the MUM-T Mission Planner (MMP). As stated in the preceding chapters the MMP follows the concept of a mixed-initiative planning assistance. Therefore, it needs plan reasoning capabilities that include first of all the ability to generate plans. This leads to the necessity of having a world model that is sufficient to plan MUM-T missions. The aspects of temporality and cost minimization are important to fulfill the requirements stated in chapter 2. Hence, the MUM-T world model of the MMP has to include a conception of time and of costs. In the case of the support through an adequate reasoning machine, such a world model automatically allows the following *interactive planning functionalities*: 1. feasibility checking of a human plan, 2.association of temporal information to human-planned tasks, 3.completeness checking of a human plan, 4.completion of a partial human plan, 5.schedule checking of a human plan, 6.re-scheduling of a human plan, 7.optimality checking of human plan, 8.generation of a complete machine plan, and 9.generation of multiple alternate complete machine plans. In order to achieve these abilities, the MMP holds domain knowledge about possible world states and allowed operations (e.g. UAV orders), which transform the world state in a specific way. In order to structure the world similarly as humans do, there should also exist object classes (types) that can be used in describing the world states and the parameters of operations (e.g. a transit operation has an aircraft and two location objects as parameters). The MMP is thereby enabled to receive an as-is state (the tactical situation), a goal state (the mission order) as well as further limiting constraints and to generate a series of operations that are associated with times and costs, i.e. an evaluable plan. The mentioned constraints can be either fixed (included in the mission order) or flexibly generated by the human operator. But it is important to note that the operator enters his constraints exclusively via the task-based UAV guidance interface explained in chapter 2. That means the orders he/she gives to the UAVs are constraining the MMP in its planning solutions, allowing it to reason upon the operator-given constraints (UAV orders). This approach, the design of the MMP and some implementation details of the working MMP prototype are described briefly in [4]. In fact, the concept described here is in accordance with the classical artificial intelligence planning approach which is why we are working with the Planning Domain Definition Language (*PDDL*) Version 2.2 [13].

5 Conclusions and Future Work

When following the Cooperative Automation approach and the CSM3 scheme the questions that remain open are how the dialog between human and machine should take place. We are trying to keep the dialog interface very basic, i.e. in the current implementation the machine instantiates the dialog with a speech synthesis output and a message box in the task-based guidance graphical user interface. Where appropriate the message box includes one or two buttons that allow the operator to invoke further aid through the assistance system or to either accept or reject its proposals respectively. In accordance with Onken and Schulte's assistant system paradigm [3] the cooperative machine is as passive as possible. Therefore, we did not implement the use case that a human can initiate a dialog with the assistant system. Instead the operator actively interacts with the OSMs in a supervisory control manner (see Fig. 2). The assistant system will initiate a monolog or dialog as soon as it detects a problem concerning the supervisory control task.

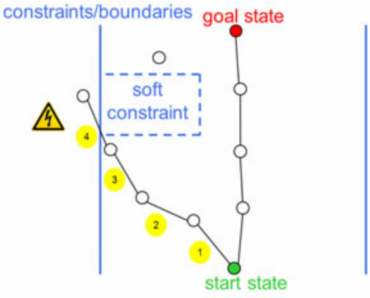


Fig. 5. Possibilities of assistance in the Cooperative Automation approach

What remains an interesting question is when an assistant system should initiate a monolog or dialog. To explain some of the different possibilities we display in Fig. 5 two planning paths inside a problem space with different world states (small circles) and hard constraints that must not be violated (blue lines) and soft constraints that should not be violated, otherwise resulting in extra costs (dotted blue lines). Although this example looks like a route planning problem it should be seen as a generic problem space visualization that is not available to the operator at the moment of planning (e.g. because of high dimensionality or dynamic problem space construction). One planning path leads quite directly to the goal state (shortest path, therefore optimal solution). Now we can imagine the operator following the other path. After each step he/she is able to leave his path and switch over to the optimal path. When should the machine tell the human that it thinks something is going wrong? After step 4 it is already too late (hard constraint violated). After step 3 the situation can still be solved but it is connected to either high path costs or soft constraint violation costs. Now the machine could also come to the conclusion that something is going wrong at step 2 or even 1, although these are valid operator actions as far as the hard constraints are concerned. But they clearly are not optimal in the view of the machine. It is important to note that the machine not necessarily has the right view. This *brittleness* is due to its limited world model. A machine that warns early and is often wrong will certainly not be accepted as a virtual teammate and will not help increasing overall system performance, because of nuisance alerts. The same counts for a machine that is often right but warns very late. Therefore, we want to examine different configurations of brittleness and interaction timing in our future experiments.

References

1. Schulte, A., Meitinger, C.: Cognitive and Cooperative Automation for Manned-unmanned Teaming Missions. In: NATO RTO-EN-SCI-208 on Advanced Automation Issues for Supervisory Control in Manned-unmanned Teaming Missions (2009)
2. Uhrmann, J., Strenzke, R., Rauschert, A., Meitinger, C., Schulte, A.: Manned-unmanned teaming: Artificial cognition applied to multiple UAV guidance. In: NATO RTO SCI-202 Symposium on Intelligent Uninhabited Vehicle Guidance Systems, Neubiberg (2009)
3. Onken, R., Schulte, A.: System-ergonomic Design of Cognitive Automation in Work Systems. Springer, Heidelberg (2010)
4. Uhrmann, J., Strenzke, R., Schulte, A.: Task-based Guidance of Multiple Detached Unmanned Sensor Platforms in Military Helicopter Operations. In: COGIS, Crawley (2010)
5. Price, H.E.: The allocation of functions in systems. Human Factors 27, 33–45 (1985)
6. Hancock, P.A., Scallen, S.F.: The future of function allocation. Ergonomics in Design 4 (1996)
7. Sheridan, T.B.: Telerobotics, Automation and Human Supervisory Control. MIT Press, Cambridge (1992)
8. Dekker, S.W.A., Woods, D.D.: MABA-MABA or Abracadabra? In: Cognition, Technology & Work. Springer, London (2002)
9. Hoyos, C.: Menschliches Handeln in technischen Systemen. In: Hoyos, C., Zimolong, B. (eds.) Enzyklopädie der Psychologie, Band D III 2, Ingenieurpsychologie, Hogrefe, Göttingen (1990)
10. Tecuci, G., Boicu, M., Cox, M.T.: Seven Aspects of Mixed-Initiative Reasoning: An Introduction to this Special Issue on Mixed-Initiative Assistants. AI Magazine 28(2) (2007)
11. Bresina, J.L., Jónsson, A.K., Morris, P.H., Rajan, K.: Mixed-Initiative Planning in MAPGEN: Capabilities and Shortcomings. In: Proceedings of ICAPS 2005 (2005)
12. Ferguson, G., Allen, J.: TRIPS: An Intelligent Integrated Problem-Solving Assistant. In: Proceedings of AAAI-1998, Madison (1998)
13. Edelkamp, S., Hoffmann, J.: PDDL2.2: The Language for the Classical Part of the 4th International Planning Competition. Technical Report 195, Albert-Ludwigs-Universität Freiburg, Institut für Informatik (2004)

Exploring the Relationship among Dimensions of Flight Comprehensive Capabilities Based on SEM

Ruishan Sun and Yang Li

Research Institute of Civil Aviation Safety, Civil Aviation University of China,
Tianjin, 300300, China
sunrsh@hotmail.com

Abstract. The structural equation model is constructed to explore the relationship among basic cognitive abilities, personality traits and mental health of pilots. A framework of hypotheses is established to test the relationship among the three dimensions based on theories in the literature. Data is gathered from 65 pilots using 3 questionnaires. The model shows that both personality traits and mental health can affect the cognitive functions significantly, and that emotion characteristic, character traits and working attitude will also have an impact on a pilot's basic cognitive ability by affecting his mental health state. The results suggest that not only flight cognitive abilities but also personality traits and mental health can affect a pilot in terms of flight performance.

Keywords: flight comprehensive capability, flight aptitude, personality traits, mental health, SEM.

1 Introduction

Since the early days of aviation history, cognitive and psychomotor skills have been seen as important aspects of being a good pilot. Moreover, personality traits of pilots have also attracted many researchers' attention. Research strategies such as observation of pilots and participant observation have been used to examine which personality traits are important in pilot selection. In particular, dated back to 1921, Dockeray and Isaacs concluded in their study on psychology in aviation that "quiet methodical men were among the best flyers" because of their "power and quick adjustment to a new situation and good judgment" [1]. In 1950s, Saul Sell and his colleagues led a project in the United States to find suitable personality measure for pilot selection. They evaluated a total of 26 personality measures and found that personality tests were better predictors of long-term criteria compared to ability tests[2][3]. With the introduction of computers in testing, a number of new personality-related concepts have also been evaluated on pilots, including measures of risk taking, assertiveness, field dependency, and attitudes [4]. Mental health is another topic that has so far been emphasized in some degrees during the evaluation process on pilots. To meet social demands or solve work-related tasks, the individual relies on different sets of resources, including knowledge, experience, and personal attributes [5]. Some theories describe mental problems as the

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 618–627, 2011.

result of factors or elements that have a negative impact on the individual. Other theories are concerned with stress, negative emotion and physical reactions.

However, few studies on the relationship among cognitive abilities, personality traits and mental health in aviation can be found in the literature. This paper tries to fill this gap and proposes to study the relationship among the three dimensions based on the structural equation model (SEM).

2 Theory and Hypotheses

2.1 Research Structure of Flight Comprehensive Capabilities

The concept of flight comprehensive capabilities is the core of flight abilities. It can be defined broadly as every internal factor that contributes to consistent behavior in different situations or, narrowly, as encompassing only cognitive functions, personality, emotions and motivation. In this study we build up a system of flight comprehensive capabilities with three dimensions including basic flight cognitive ability, personality traits and mental health. Basic flight cognitive ability is defined as a cognitive process of acquiring, processing, memorizing and applying information, including attention, working memory, judgment and decision making, speech comprehension, logical reasoning and spatial cognitive ability. Personality traits are analyzed by a matrix of character traits, emotion traits, intellect and working attitude. Character traits are personal notions that describe how a person appears to others. Emotion traits are described as emotional stability and mental experience. Intellect is a capacity for rational thought or inference or discrimination, and a willingness to pursue and explore the unfamiliar experience. Working attitude reflects individual achievement motivation, responsibility sense and self-discipline. Mental health can be defined as the psychological state of a pilot who is functioning at a satisfactory level of emotional and behavioral adjustment, which reflects his emotion, level of stress and adaptability. This paper will discuss mental health from these three aspects.

2.2 Theory and Hypotheses

A. Relationship between personality traits and flight cognitive ability. Personality traits play an important role in explaining a pilot's behavior. Pilots with certain personality characteristics such as achievement motivation and emotional stability have better flight performance. Early studies also showed that intellect reflects an individual's level of cognitive abilities, and therefore is very important. Moreover, experimental results showed that those who had the most positive attitudes towards their job performed slightly better [6]. Base on these arguments, we establish several hypotheses on the relationship between personality traits and flight cognitive abilities, as shown in Table 1.

B. Relationship between mental health and flight cognitive abilities. Mental health has a direct effect on an individual's job performance. A critical stress situation may arise when something unusual takes place during a flight. Examples of stress situations include an indication of technical difficulties or a rapidly deteriorating weather condition. Moreover, Bad interpersonal relationship or work-life conflicts may also result in a negative emotion. It is important to analyze typical reactions in such situations, be aware of how the crew responds to stress and negative emotion, and understand how it affects decisions made during the flight. According to the review provided by Orasanu [7], mental issues such as stress, depression, anxiety and so on, may have the following effects: (1) people make more errors; (2) attention is reduced; (3) working memory is reduced; (4) change of strategy: speed gains preference to accuracy. Thus, cognitive functions are subject to a number of stress-related consequences in terms of how we perceive our surroundings, process information, and make decisions. Table 2 shows hypotheses that we would like to test on the relationship between mental health and flight cognitive abilities.

C. Relationship between personality traits and mental health. There is little doubt that some working environments are generally considered stressful. However, people with certain personality characteristics experience stress or negative emotion more often and more intensely, or have stronger adaptability than others. Studies have shown that many organizations tend to reward certain types of workers, who are characterized by irritability, aggression, hostility, and ambition, because these factors usually indicate a personality of achievement striving. However, on the other hand, these personality characteristics may cause emotional problems for the individual and his or her surroundings such as stress and burnout. Furthermore, people with active coping techniques (i.e., who act strategically to handle difficult situations) generally score lower on burnout than people who use more emotion-focused coping techniques (e.g., the person attempts to deal with emotions by seeking comfort in other people) [5]. Earlier studies have found that stress can be triggered also by high interpersonal demands; while later studies have shown that individual difference is an important factor as well. Based on these studies, we build up the hypotheses on the relationship between personality traits and mental health, as shown in Table 3.

Table 1. Hypotheses on the relationship between personality traits and flight cognitive ability

H#	Hypotheses
H1	Intellect has a direct positive relationship with spatial imagination & image-thinking
H2	Intellect has a direct positive relationship with speech comprehension
H3	Intellect has a direct positive relationship with spatial cognitive & logical reasoning
H4	Intellect has a direct positive relationship with spatial perception ability
H5	Intellect has a direct positive relationship with judgment & decision making
H6	Intellect has a direct positive relationship with working memory & attention.
H7	Character traits has a direct positive relationship with speech comprehension
H8	Character traits has a direct positive relationship with judgment & decision making
H9	Working attitude has a direct positive relationship with spatial imagination
H10	Working attitude has a direct positive relationship with speech comprehension

Table 2. Hypotheses on the relationship between mental health and flight cognitive ability

H#	Hypotheses
H11	Adaptability has a direct negative relationship with spatial imagination & image-thinking
H12	Adaptability has a direct negative relationship with speech comprehension
H13	Adaptability has a direct negative relationship with spatial cognitive & logical reasoning
H14	Adaptability has a direct negative relationship with space perception ability
H15	Adaptability has a direct negative relationship with judgment & decision making
H16	Adaptability has a direct negative relationship with working memory and attention
H17	Stress situation has a direct negative relationship with spatial imagination & image-thinking
H18	Stress situation has a direct negative relationship with spatial cognitive & logical reasoning
H19	Stress situation has a direct negative relationship with working memory & attention
H20	Emotion health has a direct negative relationship with spatial imagination & image-thinking
H21	Emotion health has a direct negative relationship with speech comprehension
H22	Emotion health has a direct negative relationship with judgment & decision making
H23	Emotion health has a direct negative relationship with working memory & attention

Table 3. Hypotheses on the relationship between personality traits and mental health

H#	Hypotheses
H24	Emotion characteristics has a direct negative relationship with adaptability
H25	Character traits has a direct negative relationship with adaptability
H26	Working attitude has a direct negative relationship with adaptability
H27	Emotion characteristics has a direct negative relationship with emotional health
H28	Working attitude has a direct negative relationship with stress situation

As a summary, Fig.1 presents the theoretical framework of all the above hypotheses on the relationship among flight cognitive abilities, personality traits, and mental health.

3 Data Gathering

The structural equation model (SEM) is a statistical technique for testing and estimating causal relations using a combination of statistical data and qualitative causal assumptions [8]. We distributed a packet of 3 questionnaires to 65 pilots aged 20 to 60 years old. The three questionnaires are Flight Cognitive Capability Test, the Symptom Checklist-90 (SCL-90), and the Sixteen Personality Factor Questionnaire (16PF). Then the theoretical model was established to explore the relationship among basic flight cognitive ability, personality traits and mental health. Latent variables and manifest variables in the SEM are shown in Table 4.

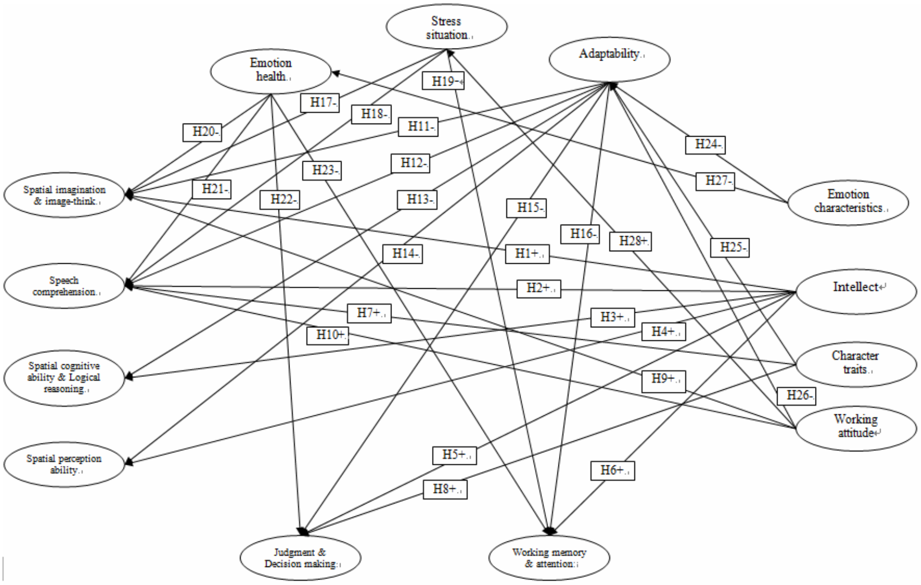


Fig. 1. Theoretical hypotheses of the study

Table 4. Latent variables and manifest variables

Latent variables	Manifest variables
Emotion characteristics(P1)	Emotional Stability, Apprehension, Liveliness, Tension
Intellect (P2)	Reasoning, Abstractedness, Openness to Change
Character traits (P3)	Warmth, Dominance, Vigilance, Privateness
Working attitude (P4)	Rule-Consciousness, Social Boldness, Sensitivity, Self-Reliance, Perfectionism
Spatial imagination & image-thinking (N1) Speech comprehension (N2) Spatial cognitive & logical reasoning (N3) Spatial perception ability (N4) Judgment & Decision making (N5) Working memory & attention (N6)	Flight cognitive Capability Test
Emotion health (E1)	Depression (DEP), Anxiety (ANX), Hostility (HOS), Phobic Anxiety (PHOB)
Stress situation (E2)	Somatization (SOM), Paranoid Ideation (PAR), Obsessive Compulsive (O-C),
Adaptability (E3)	Interpersonal Sensitivity (I-S), Psychoticism (PSY), Positive Symptom Distress Index (PSDI)

4 Data Analysis and Results

4.1 Model Modifying and Evaluation

SEM is a kind of linearity modeling tool using statistical data, which is suitable for the problem with unobservable variables especially [9, 10]. The modeling criterion is as follows. First, the smaller the ratio of chi-square and the degrees of freedom is, the better the goodness between model and practice is; second, P-value is used to test the significance level of model variables, whose value should be less than 0.1; third, the value of Comparative Fit Index (CFI） , Incremental Fit Index (IFI) and goodness of fit index(GFI) should all be greater than 0.9, while the value of Standardized Root Mean Square Residua (SRMR) should be less than 0.08.

The model is iteratively modified according to the modification indices obtained from the model itself until the estimating value is suited for the requirement of criterion. After the iterative process of testing and modifying, the goodness-of-fit statistics of the primitive model and the final revised model are shown in Table 5.

Table 5. Goodness of fit statistics

Goodness of fit statistics	primitive model	final model	Recommended values for fit
CMIN/DF	1.535	1.296	<=2
IFI	0.825	0.907	>0.9
GFI	0.656	0.696	>0.9
CFI	0.818	0.903	>0.9
RMSEA	0.091	0.068	<0.08
SRMR	0.1024	0.0955	<0.08
PGFI	0.554	0.568	>0.5
PNFI	0.559	0.601	>0.5

4.2 Results

Table 6 lists the results for the hypotheses. After deleting those invalid paths, the internal relationship will be easily found with the help of those paths showed in Fig.2.

This model is proved to be higher goodness of fit in relationship among three dimensions than the original one. The SEM model shows that both personality traits and mental health have an effect on the cognitive functions. All sub-dimensions of personality traits expect for intellect have an effect on pilots' basic cognitive abilities by affecting their mental health states. Each of the three sub-dimensions of mental health has a significant effect on the basic cognitive abilities of pilots; emotion characteristics, character traits and working attitude all have significantly direct effects on the adaptability in mental health; however, the direct influence of intellect on mental health is not significant. In addition, the intellect in personality traits has a great direct influence on the basic flight cognitive abilities. Meanwhile, character traits dramatically affect comprehension, attention and so on. Spatial sense, comprehension, attention and working memory are affected significantly by emotion health and stress level of a pilot's mental health.

Table 6. Results of hypotheses testing

H#	Relationship	P-value		Support
H1	N1<--P2	0.411		NO
H2	N2<--P2	0.011	$p<0.05$	YES
H3	N3<--P2	0.821		NO
H4	N4<--P2	0.045	$p<0.05$	NO
H5	N5<--P2	0.169		NO
H6	N6<--P2	0.748		NO
H7	N2<--P3	0.398		NO
H8	N5<--P3	0.366		NO
H9	N1<--P4	0.015	$p<0.05$	YES
H10	N2<--P4	0.008	$p<0.01$	YES
H11	N1<--E3	0.813		NO
H12	N2<--E3	0.063	$p<0.1$	NO
H13	N3<--E3	0.795		NO
H14	N4<--E3	0.009	$p<0.01$	YES
H15	N5<--E3	0.038	$p<0.05$	NO
H16	N6<--E3	0.778		NO
H17	N1<--E2	0.061	$p<0.1$	YES
H18	N3<--E2	0.801		NO
H19	N6<--E2	0.739		NO
H20	N1<--E1	0.036	$p<0.05$	YES
H21	N2<--E1	0.046	$p<0.05$	YES
H22	N5<--E1	0.033	$p<0.05$	YES
H23	N6<--E1	***	$p<0.001$	YES
H24	E3<--P1	0.009	$p<0.01$	NO
H25	E3<--P3	0.01	$p<0.05$	YES
H26	E3<--P4	0.001	$p<0.01$	NO
H27	E1<--P1	0.443		NO
H28	E2<--P4	0.768		NO

*** stand for significant difference ($p<0.001$).

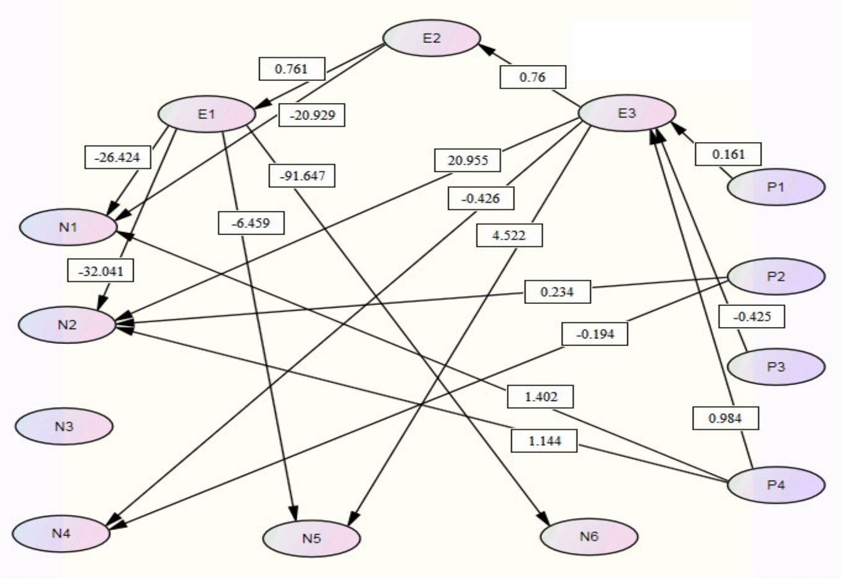


Fig. 2. Structural equation model of the study

4.3 Discussions

Theoretically, cognitive functions should be affected directly by personality traits. The model has shown that emotion characteristics, character traits and working attitude have effects on a pilot's basic cognitive ability through the way of affecting his mental health state. Emotion characteristics are recognized as emotional stability and the trend of mental experiences. Pilots in an instability mode tend to experience tension, worry and a lack of safety more often and to have mental issues more easily than those in a stability mode. According to Eysenck's awakening theory [11, 12], individuals have a high level of arousal to get the motivation to finish tasks in an emergency, which makes anxiety and stress more easily and affects cognitive functions further. Character traits are important for the development of positive emotion, good interpersonal skills and avoidance of mental problems. Past studies have found that most of the pilots have the extroversion personality [13, 14]. Researchers have become much aware of the importance of effective communication in multi-operator environments to support team performance [15]. In a high-risk situation, crew interactions with exact information and consentient situation perception can generate better decisions to solve problem compared to that of any individual. Working attitude not only directly affects spatial cognitive and speech ability, but also has impacts on a pilot's basic cognitive functions through the way of affecting his mental health. Pilots who have the most positive attitudes towards their job perform slightly better due to positive emotion and good adaptability. Meanwhile, intellect directly affects almost every aspect of cognitive functions. In particular, intellect has a more significant influence on speech and spatial ability than judgment and decision making. Furthermore, intellect has a more significant relationship with crystallized intelligence than fluid intelligence.

Moreover, it is necessary to add some more paths to the model according to the modifying results. These newly added paths further explain the relationship among stress, emotion health and adaptability. Persons with good adaptability can deal with stress problems on their merits. These people are characterized by self-assurance, extroversion, and optimism. It is easy for negative emotions to appear when something unusual takes place during their work or life. It has been proven that high workload and high risk of flight work lead pilots to experience stress more often and more intensely than people holding other jobs. At the same time, a negative emotion arises from extreme stress, and it could further affect various cognitive functions.

5 Conclusions

This study has proven again that not only flight cognitive abilities have impacts on pilots in terms of their flight performances, but also personality traits and mental health do so. In particular, interpersonal relationships, stress, emotional stability etc. have significant influences on a pilot's cognitive performance. Moreover, the relationship model has also shown that personality traits have an increasing effect on the cognitive operations with the increase of task difficulty and stress. Meanwhile, factors such as individual anxiety, depression and impulsivity also affect a pilot's flight performance. Furthermore, the research provides a tool to explore the relationship between pilots and their operational traits via QAR (Quick Access Recorder) data, which is useful for managing pilots' human errors.

References

1. Dockeray, F.C., Isaacs, S.: Psychological research in aviation in Italy, France, England, and the American Expeditionary Forces. Journal of Comparative Psychology 1, 115–148 (1921)
2. Sells, S.B.: Development of a personality test battery for psychiatric screening of flying personnel. Journal of Aviation Medicine 26, 35–45 (1955)
3. Sells, S.B.: Further developments on adaptability screening for flying personnel. Aviation Medicine 27, 440–451 (1956)
4. Hunter, D.R., Burke, E.F.: Handbook of pilot selection. Avebury Aviation, Aldershot (1995)
5. Hunter, D.R., Martinussen, M.: Aviation Psychology and Human Factor. CRC Press, Boca Raton (2010)
6. Lang-Ree, O.C., Martinussen, M.: Applicant reactions and attitudes towards the selection procedure in the Norwegian Air Force. Human Factors and Aerospace Safety 6, 345–358 (2006)
7. Orasanu, J.: Stress and naturalistic decision making: Strengthening the weak links. Ashgate, Aldershot (1997)
8. Simon, H.: Causal ordering and identifiability. In: Hood, W.C., Koopmans, T.C. (eds.) Studies in Econometric Method, pp. 49–74. Wiley, New York (1953)
9. Boucard, A., Marchand, A., Noguès, X.: Reliability and validity of structural equation modeling applied to neuroimaging data: A simulation study. Journal of Neuroscience Methods 166(2), 278–292
10. Jietai, H., Zhonglin, W., Zijun, C.: Structural equation model and its application. Educational Science Publishing House, Beijing (2005)

11. Eysenek, H.J., Eysenek, M.W.: Personality and individual differences. Plenum, NewYork (1985)
12. Eysenck, H.J.: Relation between intelligence and personality. Perceptual and Motor Skills 32, 637–638 (1971)
13. Xiaowei: The research on the personality traits influencing flying cadets emotional stability. Department of Psychology Faculty of Aerospace Medicine, The Fourth Military Medical University, Xi'an (2002)
14. Yu, H., Jiao, Z.: A study of personality characteristics in pilots in the civil aviation. Shandong Arch. Psychiatry 19(2), 113–115 (2006)
15. Harris, D., Muir, H.C.: Contemporary issues in human factors and aviation safety. United Kingdom, Surrey, Ashgate (2005)
16. Eysenck, H.J.: Personality and intelligence: psychometric and experimental approaches. In: Sternberg, R.J. (ed.) Personality and Intelligence, pp. 3–31. Cambridge University, New York (1994)
17. McCann, S.J.H.: Emotional Health and the Big Five Personality Factors at the American State Level. Cape Breton University, Sydney
18. Liu, Y., Gordon-Becker, S.E.: An introduction to human factors engineering. Pearson Education, New Jersey (2004)
19. Ellis, B.B., Mead, A.D.: Assessment of the Measurement Equivalence of a Spanish Translation of the 16PF Questionnaire. In: Educational and Psychological Measurement, vol. 60(5), pp. 787–807. Sage Publications, Inc., Thousand Oaks (2004)
20. Chernyshenko, O.S., Stark, S., Chan, K.Y.: Investigating the Hierarchical Factor Structure of the Fifth Edition of the 16PF: an Application of the Schmid-Leiman Orthogonalization Procedure. In: Educational and Psychological Measurement, vol. 61(2), pp. 290–302. Sage Publications, Inc., Thousand Oaks (2001)
21. Hu, S.-J., Jou, S.-Y., Liu, Y.-H.: Structural equation model for brand image measurement of Jeans. In: Ninth International Conference on Hybrid Intelligent Systems, pp. 85–94 (2009)
22. Liu, T.-z., Li , Z.-x.: Structural Equation Model for the Affecting Factors of Safety management Capability of Coal Mine. In: 2008 International Workshop on Modeling, Simulation and Optimization, pp. 74-77 (2008)

A Generic After Action Review Capability for Game-Based Training

Casey L. Thurston and Glenn A. Martin

University of Central Florida Institute for Simulation and Training,
3100 Technology Pkwy, Orlando Fl 32826
{cthurston,martin}@ist.ucf.edu

Abstract. Recent years have seen a surge of interest in game-based training by the military. Game-based simulation possesses a number of potential benefits including decreased testbed complexity and cost, increased agility regarding both software and hardware, and the possibility of increased effectiveness relative to traditional training methods. One area of weakness in game-based training is the difficulty in supporting an after action review (AAR). The paper explores the emerging problems faced by systems attempting to facilitate AAR in game-based training scenarios. It presents an architecture that addresses or circumvents these issues in a flexible and game-agnostic manner, and details the limitations introduced by such an approach. A discussion on future work leveraging the plugin-based SOCRATES architecture to augment video for improved training is included.

Keywords: after action review, AAR, game-based training.

1 Background

An after action review (AAR) is the process by which trainees review the events of a training exercise and identify ways to improve future performance. AAR software for training simulators often relies on some level of integration with the simulator and its content, reading network traffic and graphical assets in order to present a visual representation of events to the user. Detailed statistics can be compiled into useful metrics related to general or specific training goals. A flexible 3D view may show the user the action from any angle at any time, and augmented elements such as entity labels or path lines may be inserted into the scene to assist the instructor. These features have been shown to facilitate improved training in a virtual setting [1, 3].

The Army has leveraged AAR for improved training for many years [2]. We have previously examined AAR via the Dismounted Infantry Virtual After Action Review System (DIVAARS) [3]. DIVAARS allows an AAR Facilitator to view the scene in 3D, select desired augmented elements, display statistical tables, bookmark events in time and jump to these bookmarks easily during review, and draw lines and figures over top of the virtual scene to assist training (telestration). DIVAARS has been well received by Soldiers in representing the events of a scenario and supporting discussion towards training [1].

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2011, LNAI 6781, pp. 628–634, 2011.

2 Game-Based Simulation

Game-based simulation has entered a period of scrutiny for its potential to fulfill military training objectives [4]. In this section we will briefly discuss attributes of game-based training, motivations behind this shift in focus, and the implications for AAR.

Game-based simulation involves participants with either video game console systems or personal computers and common peripherals (monitor, keyboard, and mouse most often) running game software towards a research or training objective. The game may be written with training domain-specific tasks in mind (Virtual Battlespace 2: Army from Bohemia Interactive, Inc), or it may be adapted from commercial software via scripting or program code modification (Game-Distributed Interactive Simulation from Research Networks, Inc). Multiple stations may be networked for team training exercises, in a single room or widely distributed.

While training tasks may be similar, game-based simulators differ from their immersive counterparts in a number of ways. One such compelling difference is cost. Early immersive simulators represented a significant investment in both user interface hardware (head-mounted displays, tracking systems, etc...) and in specialized computers to perform simulation logic and produce the resulting graphics. Personal computers of sufficient power are now available at a substantial reduction in cost, with the result that interface devices now constitute a much larger proportion (if not the majority) of the total investment [4]. That overall cost could be mitigated drastically if comparable training could be achieved using modern PCs and common peripherals.

Other motivations include testbed agility and user familiarity. Because games are designed for compatibility with a broad spectrum of hardware, upgrading to new hardware or switching to a new game is often as simple as running one or more installers and performing basic configuration. The increased accessibility of modern PCs may also correspond to an increase in pre-existing familiarity with games and their interfaces among Soldiers, which could in turn result in increased transfer of skills[1].

However, with these potential strengths come certain weaknesses, especially regarding after action review. While AAR software can often be integrated with standards-based simulators with ease, the work required to interface with a game-based simulator may vary drastically. As commercial products, games are not required to conform to simulation standards (such as DIS or HLA). Many obfuscate their network traffic or employ unpublished packet formats. Some encrypt their graphical assets, making synchronization of visual elements between the original training scenario and the AAR more difficult and costly. These details differ from game to game, so maintaining compatibility between the game software and its AAR can require continuous effort. In order to retain the simplicity and flexibility of a game-based training testbed while still facilitating training via AAR, a more generic solution for AAR against games is desired.

[1] It should be noted that the authors are not aware of any major study which demonstrates an increase in familiarity with games over time with specific regard to Soldiers.

3 GEAARS

Our exploration into an AAR capability for game-based simulation resulted in the Generic Engine for After Action Review Scenarios (GEAARS). This section enumerates our requirements, discusses obstacles and the means by which they were overcome, and describes the capabilities and limitations of the system.

Our requirements for GEAARS reside at the intersection of those capabilities offered by DIVAARS as enumerated above and those which were deemed satisfiable in a game-agnostic manner. With the goal of facilitating AAR against games which neither expose their databases nor conform to simulation networking standards (DIS or HLA), we focused on the capture and use of those data which are available in the overwhelming majority of cases: the frames of video presented to the trainee.

In order for GEAARS to facilitate an AAR it must capture video feeds from each of the trainee stations, store them for later playback, and consolidate the data on a single Facilitator station. Additional requirements derived from DIVAARS include the following: recording should be a stealth operation with little or no impact on the exercise; the time between the end of the exercise and the ability to begin an AAR should be kept under 10 minutes; a Facilitator should be able to monitor the exercise as it takes place; a distributed exercise and review involving trainees in different rooms or on different continents should be possible; an AAR Facilitator must be able to navigate freely through time during replay, bookmark critical events, and draw figures to assist comprehension and training (telestration).

3.1 Gathering Data

We focus first on the problem of recording the necessary data and making it available for AAR. The client/server relationship of the trainee and AAR Facilitator stations suggests two general approaches, distinguished by the order of the video capture and consolidation operations.

In the consolidate-then-capture approach, the raw video feed of each trainee machine is streamed via a secondary output on its video card across video cable (DVI, HDMI, or similar) to the facilitator station. These video streams are then captured and stored in files directly at the master station where they may be utilized during an after action review without further processing. Because video is streamed in real-time for capture purposes, monitoring the exercise is as simple as presenting incoming video feeds to the user. Down-time between an exercise and its AAR is nominal. Despite these benefits, the transmission and capture of multiple feeds of raw video presents two significant logistical issues. The bandwidth requirements of transmitting uncompressed video long distances are often prohibitive with regards to the distributed case, and capturing of those feeds on the destination machine requires input and throughput capabilities exceeding those of even modern high-end machines.

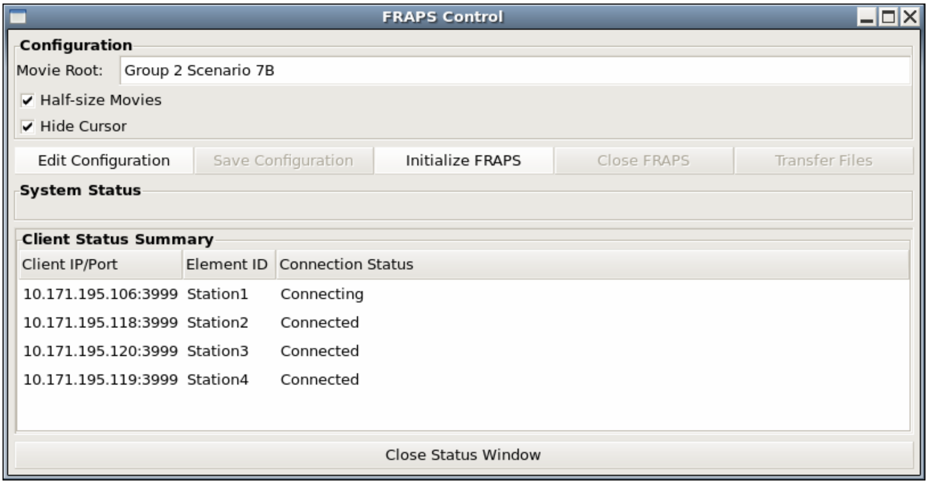
In the capture-then-consolidate approach, video is captured and stored on each of the trainee machines during the exercise. The resulting video files are transferred to the master station at some point before AAR is to be performed. The down-time between exercise and AAR is dependent on the consolidation of those files on the master station, a task whose time and bandwidth requirements scale with the number of participants and the length of the exercise. As both capture and transfer operations must be unnoticeable by the trainee, network latency-inducing file transfer must wait until after the exercise has concluded. Raw video at full quality is sufficiently large relative to the available 10 minute window that the bandwidth requirements again become prohibitive in the distributed case.

To satisfy the breadth of constraints, GEAARS employs a hybrid solution supplementing the capture mechanism of the second approach with a multi-stage transfer process. A low-profile commercial application called FRAPS performs capture of video on each of the trainee machines [5]. To provide a degree of monitoring during the exercise, the GEAARS trainee client captures screenshots at a fixed interval and spends a minimal portion of bandwidth streaming them to the Facilitator station. Upon conclusion of the exercise, the GEAARS client uses the FFMPEG video library to compress captured video files prior to their transfer [6]. We leverage the multi-core architecture of modern computers to achieve transfer of many times as much compressed video as that which is possible uncompressed (see Table 1, transfer times for 10, 30, and 90 minutes with no compression, single-process, and multi-process demonstrating impact of parallel compression). The target compression quality is configurable on a per-trainee basis in order to accommodate differences in computing power and available bandwidth (differences which are most prevalent in the massively distributed case, where some trainees may be in the same room as the Facilitator and others may reside on a different continent). After the compressed video is transferred to satisfy the needs of immediate AAR, the raw uncompressed footage may be transferred for use in later instruction. This approach achieves the primary objectives in a manner that requires no additional hardware, scales to a large number of participants, and supports distributed exercises.

Table 1. Times for consolidating video by each of three methods. All values are times in MM:SS format. Times for compressed video include both compression and transfer. [1]Note that the 10-minute exercise is stored in a single file and thus cannot benefit from parallel compression.

Length of exercise	Uncompressed	Compressed (single core)	Compressed (quad core)
10:00	3:49	5:29	5:29[1]
30:00	11:40	13:08	8:35
90:00	35:22	25:11	9:47

The GEAARS Facilitator station, an application built upon the plugin-based SOCRATES AAR engine, is responsible for coordinating the GEAARS trainee clients and their data [7]. At the start of an exercise it connects to each specified client, asserts the desired capture and compression configuration options, and responds to a click of the record button by issuing a synchronized message initiating capture. It monitors the network status of each client and will attempt to resume where it left off in the case of a lost network connection. When the exercise has concluded, the user clicks a stop button to cease capture followed by the transfer button to begin the compression and consolidation process. The AAR may commence as soon as transfer of video to the GEAARS Facilitator station has finished.



Movie Root: Group 2 Scenario 7B

☑ Half-size Movies

☑ Hide Cursor

Client IP/Port	Element ID	Connection Status
10.171.195.106:3999	Station1	Connecting
10.171.195.118:3999	Station2	Connected
10.171.195.120:3999	Station3	Connected
10.171.195.119:3999	Station4	Connected

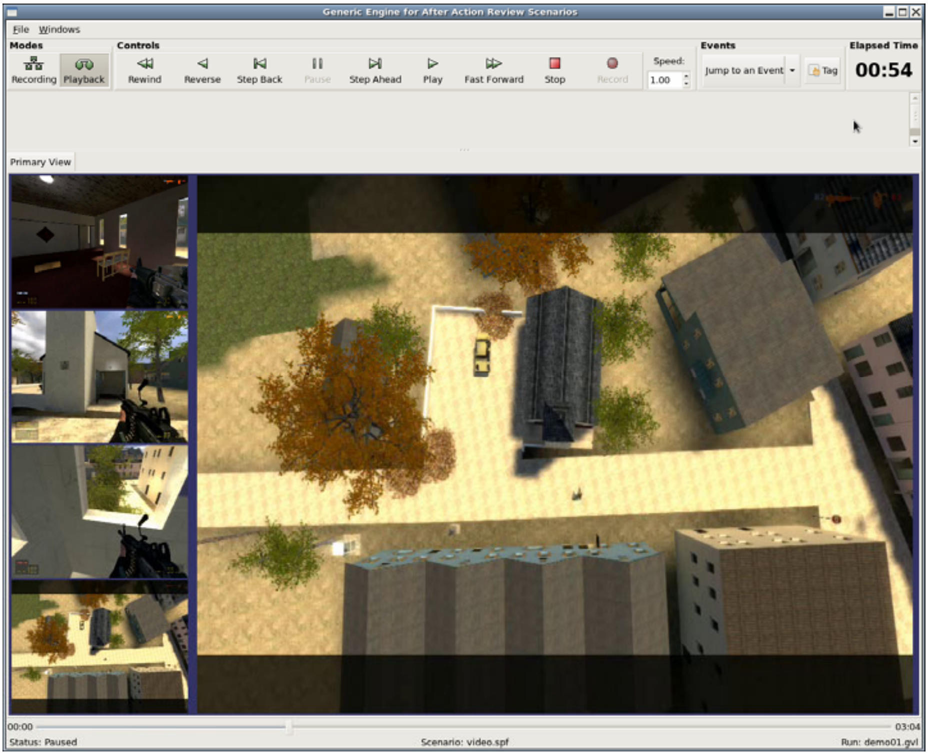
Fig. 1. The FRAPS Control screen provides the GEAARS Facilitator with configuration, control, and monitoring capabilities

3.2 After Action Review

GEAARS attempts to preserve the replay capabilities of existing virtual AAR software. Where DIVAARS features a single flexible 3D view, GEAARS allows the Facilitator to present video in many different ways: a single viewpoint, a split display featuring all viewpoints equally, one or two featured viewpoints with thumbnail or preview images of the others, etc... The ability to bookmark key events and jump freely through time is fully supported, as is the telestration capability. A trainee station in stealth mode has even been used to provide a limited map capability by placing it in an overhead view.

In the special case where the network traffic is available, GEAARS can leverage entity position and orientation data to augment video sequences with such instructional elements as element name tags or paths. However, augmentations in this limited mode always appear as an overlay to the video, with the consequence that they are visible even when one might desire them to be obscured (the name tag of a player behind a building, for example).

Certain features desirable for virtual AAR systems remain lacking in this version of GEAARS. The range of perspectives offered by the freely-moving camera of DIVAARS is notably absent; only the perspectives experienced by trainee stations are available during the AAR. A related missing feature is the DIVAARS building scalpel, which allows the user to peel off the floors of a building to view events that occur within it. Without an understanding of the underlying geometry, this valuable feature remains a practical impossibility.



Status: Paused

Scenario: video.spf

Run: demo01.gvl

Fig. 2. The GEAARS interface in a single-column preview configuration. The viewpoints of other trainees are available on the left, while the maximized stealth entity map view demonstrates a Soldier crossing an open road between buildings.

3.3 Testbed

Our testbed consists of nine trainee stations and one Facilitator station. The trainee stations are capable of running the Game-Distributed Interactive Simulation (G-DIS) system from Research Networks, Inc., the On-Line Interactive Virtual Environment (OLIVE) system from SAIC, Inc., or the Virtual Battlespace 2: Army (VBS2: Army) system from Bohemia Interactive, Inc. Stations communicate across a gigabit local area network. GEAARS captures data during exercises using any of these systems without any code or configuration changes.

4 Conclusions and Future Work

Game-based simulation and training shows no signs of a decreased presence in coming years. In order to ascertain the merits of game-based simulation and fully leverage it for training we must achieve an AAR capability comparable to that of standards-based simulators. GEAARS provides a foundation for this capability, but still lacks features from which AAR for game-based training may benefit.

Ongoing research in the fields of augmented reality and computer vision may offer solutions to some of these needs while maintaining the flexibility desired of a game-agnostic AAR system. Recent work in 3D scene reconstruction from uncalibrated video and photographs may prove especially adaptable, furnishing databases for AAR without calibration or time-consuming off-line modeling [8, 9]. By image-matching trainee video to these calculated databases over the course of an exercise, it may be possible to estimate the position and orientation of participants on a per-frame basis. Further leveraging machine learning, keystroke tracking, and techniques from the field of computer vision may allow even greater detail to be extracted for use in game-agnostic AAR.

Acknowledgements. The authors would like to acknowledge the support of the Army Research Institute (ARI) for the funding of this work. Opinions expressed are those of the authors and do not necessarily represent an official position of the Department of the Army or the Army Research Institute.

References

1. Lampton, D.R., Bliss, J.P., Martin, G.A.: Performance Measurement and Training Feedback in a Military Collaborative Virtual Environment. In: HCI International (2005)
2. Morrison, J.E., Meliza, L.L.: Foundations of the After Action Review Process. Special Report 42, United States Army Research Institute for the Behavioural and Social Sciences (1999)
3. Knerr, B.W., Lampton, D.R., Martin, G.A., Washburn, D.A., Cope, D.: Developing an After Action Review System for Virtual Dismounted Infantry Simulations. In: The Interservice/Industry Training, Simulation & Education Conference (2002)
4. Knerr, B.W.: Current Issues in the Use of Virtual Simulations for Dismounted Soldier Training. In: Virtual Media for Military Applications, Meeting Proceedings RTO-MP-HFM-136, Paper 21, pp. 21-1 – 21-12. RTO, Neuilly-sur-Seine, France (2006)
5. FRAPS, Real-time video capture and benchmarking, `http://www.fraps.com`
6. FFMPEG, `http://www.ffmpeg.org`
7. Martin, G.A., Daly, J., Lampton, D.R.: An After Action Review Engine for Training in Multiple Domains. In: Simulation & Education Conference on The Interservice/Industry Training (2008)
8. Frahm, J.-M., Pollefeys, M., Clipp, B., Gallup, D., Raguram, R., Wu, C., Zach, C.: 3D Reconstruction of Architectural Scenes from Unaclibrated Video Sequences. In: International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences, XXXVIII (2009)
9. Frahm, J.-M., Pollefeys, M., Lazebnik, S., Clipp, B., Gallup, D., Raguram, R., Wu, C.: Fast Robust Reconstruction of Large Scale Environments. In: 44th Annual Conference on Information Sciences and Systems (2010)

Author Index