

Pedro Campos Nicholas Graham
Joaquim Jorge Nuno Nunes
Philippe Palanque Marco Winckler (Eds.)

LNCS 6946

Human-Computer Interaction – INTERACT 2011

13th IFIP TC 13 International Conference
Lisbon, Portugal, September 2011
Proceedings, Part I

1
Part I



ifip

 Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Pedro Campos Nicholas Graham
Joaquim Jorge Nuno Nunes
Philippe Palanque Marco Winckler (Eds.)

Human-Computer Interaction – INTERACT 2011

13th IFIP TC 13 International Conference
Lisbon, Portugal, September 5-9, 2011
Proceedings, Part I

Volume Editors

Pedro Campos
Nuno Nunes
University of Madeira
9020-105, Funchal, Portugal
E-mail: {pcampos, njn}@uma.pt

Nicholas Graham
Queen's University
Kingston, ON K7L 3N6, Canada
E-mail: graham@equis.cs.queensu.ca

Joaquim Jorge
Instituto Superior Técnico
1049-001 Lisbon, Portugal
E-mail: jaj@inesc.pt

Philippe Palanque
Marco Winckler
University Paul Sabatier
31062 Toulouse Cedex 9, France
E-mail: {palanque, winckler}@irit.fr

ISSN 0302-9743
ISSN 978-3-642-23773-7
DOI 10.1007/978-3-642-23774-4
Springer Heidelberg Dordrecht London New York

e-ISSN 1611-3349
e-ISBN 978-3-642-23774-4

Library of Congress Control Number: 2011935338

CR Subject Classification (1998): H.5.2, H.5.3, H.3-5, I.2.10, D.2, K.3-4, K.8

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

© IFIP International Federation for Information Processing 2011

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

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

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

Advances in interactivity, computing power, mobile devices, large displays and ubiquitous computing offer an ever-increasing potential for empowering users. This can happen within their working environment, in their leisure time or even when extending their social skills. While such empowerment could be seen as a way of connecting people in their workspace, home or on the move, it could also generate gaps requiring larger effort and resources to fruitfully integrate disparate and heterogeneous computing systems.

The conference theme of INTERACT 2011 was “building bridges” as we believe human–computer interaction (HCI) is one the research domains more likely to significantly contribute to bridging such gaps. This theme thus recognizes the interdisciplinary and intercultural spirit that lies at the core of HCI research. The conference had the objective of attracting research that bridges disciplines, cultures and societies. Within the broad umbrella of HCI, we were in particular seeking high-quality contributions opening new and emerging HCI disciplines, bridging cultural differences, and tackling important social problems. Thus, INTERACT 2011 provided a forum for practitioners and researchers to discuss all aspects of HCI, including these challenges. The scientific contributions gathered in these proceedings clearly demonstrate the huge potential of that research domain to improving both user experience and performance of people interacting with computing devices. The conference also is as much about building bridges on the human side (between disciplines, cultures and society) as on the computing realm.

INTERACT 2011 was the 13th conference of the series, taking place 27 years after the first INTERACT held in early September 1984 in London, UK. Since INTERACT 1990 the conferences have taken place under the aegis of the UNESCO International Federation for Information Processing (IFIP) Technical Committee 13. This committee aims at developing the science and technology of the interaction between humans and computing devices through different Working Groups and Special Interests Groups, all of which, together with their officers, are listed within these proceedings.

INTERACT 2011 was the first conference of its series to be organized in cooperation with ACM SIGCHI, the Special Interest Group on Computer–Human Interaction of the Association for Computing Machinery. We believe that this cooperation was very useful in making the event both more attractive and visible to the worldwide scientific community developing research in the field of HCI.

We thank all the authors who chose INTERACT 2011 as the venue to publish their research. This was a record year for the conference in terms of submissions in the main technical categories. For the main Technical Program there were a total of 680 submissions, including 402 long and 278 short papers, out of which we accepted 171 (111 long and 60 short submissions), for a combined acceptance rate of less than 25%. Overall, from a total of 741 submissions for all tracks, 290 were accepted, as follows:

- 111 Full Research Papers
- 60 Short Research Papers
- 54 Interactive Poster Papers
- 17 Doctoral Consortium Papers
- 16 Workshops
- 12 Tutorials
- 5 Demonstrations
- 6 Organizational Overviews
- 4 Industrial Papers
- 3 Special Interest Groups
- 2 Panels

Our sincere gratitude goes to the members of our Program Committee (PC), who devoted countless hours to ensure the high quality of the INTERACT Conference. This year, we improved the reviewing process by moving to an associate chair model. With almost 700 submitted papers, it is impossible for the PC Chairs to read every paper. We recruited 103 Associate Chairs (ACs), each of whom handled up to 12 papers. The ACs recruited almost 800 external reviewers, guaranteeing that each paper was reviewed by three to six referees. ACs also provided a meta-review. Internal discussion among all the reviewers preceded the final decision between the PC Chairs and the AC. This herculean effort was only possible due to the diligent work of many people. We would like to thank you all for the effort and apologize for all the bullying required to get the work done on time.

In addition, sincere thanks must be extended to those whose contributions were essential in making it possible for the conference to happen and for these proceedings to be produced. We owe a great debt to the Conference Committees, the members of the International Program Committee and the numerous reviewers who had to review submissions from the various categories. Similarly, the members of the conference Organizing Committee, the staff at INESC-ID, especially Manuela Sado, deserve much appreciation for their tireless help with all aspects of planning and managing the many administrative and organizational issues. We would like to especially thank Tiago Guerreiro for his dedication with the Student Volunteer program, and José Coelho who worked tirelessly to make the online program a reality. Thanks are also due to Alfredo Ferreira for keeping and single-handedly maintaining the website, and to Pedro Campos and Marco Winkler for the superb work done with the conference proceedings. Finally, our thanks go to all the authors who actually did the scientific work and especially to the presenters who took the additional burden of discussing the results with their peers at INTERACT 2011 in Lisbon.

July 2011

Nicholas Graham
Daniel Gonçalves
Joaquim Jorge
Nuno Nunes
Philippe Palanque

IFIP TC13

Established in 1989, the International Federation for Information Processing Technical Committee on Human–Computer Interaction (IFIP TC13) is an international committee comprising 30 national societies and 7 working groups, representing specialists in human factors, ergonomics, cognitive science, computer science, design and related disciplines. INTERACT is its flagship conference, staged biennially in different countries in the world.

IFIP TC13 aims to develop the science and technology of human–computer interaction (HCI) by encouraging empirical research; promoting the use of knowledge and methods from the human sciences in design and evaluation of computer systems; promoting better understanding of the relation between formal design methods and system usability and acceptability; developing guidelines, models and methods by which designers may provide better human-oriented computer systems; and, cooperating with other groups, inside and outside IFIP, to promote user-orientation and humanization in system design. Thus, TC13 seeks to improve interactions between people and computers, encourage the growth of HCI research and disseminate these benefits world-wide.

The main orientation is toward users, especially the non-computer professional users, and how to improve human–computer relations. Areas of study include: the problems people have with computers; the impact on people in individual and organizational contexts; the determinants of utility, usability and acceptability; the appropriate allocation of tasks between computers and users; modelling the user to aid better system design; and harmonizing the computer to user characteristics and needs.

While the scope is thus set wide, with a tendency toward general principles rather than particular systems, it is recognized that progress will only be achieved through both general studies to advance theoretical understanding and specific studies on practical issues (e.g., interface design standards, software system consistency, documentation, appropriateness of alternative communication media, human factors guidelines for dialogue design, the problems of integrating multi-media systems to match system needs and organizational practices, etc.).

IFIP TC13 stimulates working events and activities through its working groups (WGs). WGs consist of HCI experts from many countries, who seek to expand knowledge and find solutions to HCI issues and concerns within their domains, as outlined below.

In 1999, TC13 initiated a special IFIP Award, the Brian Shackel Award, for the most outstanding contribution in the form of a refereed paper submitted to and delivered at each INTERACT. The award draws attention to the need for a comprehensive human-centered approach in the design and use of information technology in which the human and social implications have been taken into

account. Since the process to decide the award takes place after papers are submitted for publication, the award is not identified in the proceedings.

WG13.1 (Education in HCI and HCI Curricula) aims to improve HCI education at all levels of higher education, coordinate and unite efforts to develop HCI curricula and promote HCI teaching.

WG13.2 (Methodology for User-Centered System Design) aims to foster research, dissemination of information and good practice in the methodical application of HCI to software engineering.

WG13.3 (HCI and Disability) aims to make HCI designers aware of the needs of people with disabilities and encourage development of information systems and tools permitting adaptation of interfaces to specific users.

WG13.4 (also WG2.7) (User Interface Engineering) investigates the nature, concepts and construction of user interfaces for software systems, using a framework for reasoning about interactive systems and an engineering model for developing user interfaces.

WG13.5 (Human Error, Safety and System Development) seeks a framework for studying human factors relating to systems failure, develops leading-edge techniques in hazard analysis and safety engineering of computer-based systems, and guides international accreditation activities for safety-critical systems.

WG13.6 (Human-Work Interaction Design) aims at establishing relationships between extensive empirical work-domain studies and HCI design. It promotes the use of knowledge, concepts, methods and techniques that enables user studies to procure a better apprehension of the complex interplay between individual, social and organizational contexts and thereby a better understanding of how and why people work in the ways that they do.

WG13.7 (Human-Computer Interaction and Visualization) is the newest of the working groups under the TC.13. It aims to establish a study and research program that combines both scientific work and practical applications in the fields of human-computer interaction and visualization. It integrates several additional aspects of further research areas, such as scientific visualization, data mining, information design, computer graphics, cognition sciences, perception theory, or psychology, into this approach.

New WGs are formed as areas of significance to HCI arise. Further information is available on the IFIP TC13 website: <http://csmobile.upe.ac.za/ifip>

IFIP TC13 Members

Australia

Judy Hammond
Australian Computer Society

Austria

Andreas Holzinger
Austrian Computer Society

Belgium

Monique Noirhomme-Fraiture
*Federation des Associations
Informatiques de Belgique*

Brazil

Simone Diniz Junqueira Barbosa
(TC 13 secretary)
Brazilian Computer Society (SBC)

Bulgaria

Kamelia Stefanova
Bulgarian Academy of Sciences

Canada

Heather O'Brian
*Canadian Information Processing
Society*

China

Zhengjie Liu
Chinese Institute of Electronics

Cyprus

Panayiotis Zaphiris
Cyprus Computer Society

Czech Republic

Vaclav Matousek
*Czech Society for Cybernetics and
Informatics*

Denmark

Annelise Mark Pejtersen
*Danish Federation for Information
Processing*

Finland

Kari-Jouko Rähkä
*Finnish Information Processing
Association*

France

Philippe Palanque (TC 13 vice chair)
*Societe des Electriciens et
des Electroniciens (SEE)*

Germany

Tom Gross
Gesellschaft für Informatik

Hungary

Cecilia Sik Lanyi
*John v. Neumann Computer Society
(NJSZT)*

Iceland

Marta Kristin Larusdottir
*The Icelandic Society for Information
Processing (ISIP)*

India

Anirudha Joshi
Computer Society of India

Italy

Fabio Paternò
Italian Computer Society

Ireland

Liam J. Bannon

*Irish Computer Society***Japan**

Masaaki Kurosu

*Information Processing Society
of Japan***Kenya**

Daniel Orwa Ochieng

*Computer Society of Kenya***Malaysia**

Chui Yin Wong

*Malaysian National Computer
Confederation***New Zealand**

Mark Apperley

*New Zealand Computer Society
(NZCS)***Nigeria**

Chris Nwannenna

*Nigeria Computer Society***Norway**

Dag Svanes

*Norwegian Computer Society***Poland**

Juliusz L. Kulikowski

*Poland Academy of Sciences***Portugal**

Joaquim A. Jorge

*Associação Portuguesa de Informática***Singapore**

Henry Been-Lirn Duh

*Singapore Computer Society***South Africa**

Paula Kotzé

*The Computer Society of South Africa***Spain**

Julio Abascal

*Asociación de Técnicos de Informática
(ATI)***Sweden**

Jan Gulliksen TC 13 (chair)

*Swedish Interdisciplinary Society for
Human-Computer Interaction
(STIMDI) - Swedish Computer Society***Switzerland**

Ute Klotz

*Swiss Association for Research in
Information Technology
SARIT***The Netherlands**

Gerrit C. van der Veer

*Nederlands Genootschap voor
Informatica***UK**

Andrew Dearden

*British Computer Society (BCS)***USA-based**

John Karat

*Association for Computing Machinery
(ACM)*

Nahum Gershon

*The Computer Society, Institute of
Electrical & Electronics Engineers
(IEEE-CS)***Expert members**Nikos Avouris, *Greece*Paula Kotzé, *South Africa*Gitte Lindegaard, *Canada*Annelise Mark Pejtersen, *Denmark*Marco Winckler, *France*

Working Group Chairpersons

WG13.1 (Education in HCI and HCI Curricula)

Lars Oestreicher, *Sweden*

SIG13.1 (Interaction Design and International Development)

Janet Read, *UK*

WG13.2 (Methodology for User-Centered System Design)

Peter Forbrig, *Germany*

SIG13.2 (Interaction Design and Children)

Panos Markopoulos, *The Netherlands*

WG13.3 (HCI and Disability)

Gerard Weber, *Germany*

WG13.4 (joint with WG 2.7) (User Interface Engineering)

Fabio Paternó, *Italy*

WG13.5 (Human Error, Safety, and System Development)

Philippe Palanque, *France*

WG13.6 (Human-Work Interaction Design)

Torkil Clemmensen, *Denmark*

WG13.7 (Human-Computer Interaction and Visualization)

Achim Ebert, *Germany*

INTERACT 2011 Technical Committee

Conference Committee

General Co-chairs

Joaquim A. Jorge, *Portugal*

Philippe Palanque, *France*

Honorary Co-chairs

Larry Constantine, *Portugal*

Don Norman, *USA*

Anneliese Mark Pejtersen, *Denmark*

Technical Program Co-chairs

Daniel Gonçalves, *Portugal*

Nick Graham, *Canada*

Nuno Nunes, *Portugal*

Technical Program Committee

Demonstrations Co-chairs

Verónica Orvalho, *Portugal*

Greg Philips, *Canada*

Doctoral Consortium Co-chairs

Gitte Lindgaard, *Canada*

Manuel João Fonseca, *Portugal*

Full Papers Co-chairs

Nick Graham, *Canada*

Nuno Nunes, *Portugal*

Industrial Program Co-chairs

Antonio Câmara, *Portugal*

Miguel Dias, *Portugal*

Stacy Hobson, *USA*

Oscar Pastor, *Spain*

Virpi Roto, *Finland*

Interactive Posters Co-chairs

Adérito Marcos, *Portugal*

Monique Noirhomme-Fraiture, *Belgium*

Keynote Speakers Co-chairsJohn Karat, *USA*Jean Vanderdonckt, *Belgium***Organization Overviews Co-chairs**Teresa Chambel, *Portugal*Mary Czerwinski, *USA***Panels Co-chairs**Regina Bernhaupt, *Austria*Nuno Correia, *Portugal*Peter Forbrig, *Germany***Short Papers Co-chairs**Daniel Gonçalves, *Portugal***Special Interest Groups (SIGs) Co-chairs**Gerrit van der Veer, *The Netherlands*Teresa Romão, *Portugal***Student Design Competition Co-chairs**Simone Diniz Junqueira Barbosa, *Brazil*Luis Carriço, *Portugal***Tutorials Co-chairs**José Creissac Campos, *Portugal*Paula Kotzé, *South Africa***Workshops Co-chairs**Julio Abascal, *Spain*Nuno Guimarães, *Portugal***Organizing Committee****Local Organization Co-chairs**Alfredo Ferreira, *Portugal*Pauline Jepp, *Portugal*Manuela Sado, *Portugal***Multimedia Conferencing Co-chairs**José Coelho, *Portugal*Lars Oestreicher, *Sweden***Publications Co-chairs**Padro Campos, *Portugal*Marco Winckler, *France*

Publicity Co-chairs

Paula Alexandra Silva, *Portugal*
Tiago Guerreiro, *Portugal*

Student Volunteers Co-chairs

Tiago Guerreiro, *Portugal*
Xavier Ferre, *Spain*
Effie Law, *UK*

Website Co-chairs

Alfredo Ferreira, *Portugal*

Associate Chairs - Full Papers

Julio Abascal, <i>Spain</i>	Phil Gray, <i>UK</i>
Jose Abdelnour-Nocera, <i>UK</i>	Tom Gross, <i>Germany</i>
Silvia Abrahão, <i>Spain</i>	Mark D Gross, <i>USA</i>
Vincent Aleven, <i>USA</i>	Jan Gulliksen, <i>Sweden</i>
Nikolaos Avouris, <i>Greece</i>	Michael Haller, <i>Austria</i>
Cecilia Baranauskas, <i>Brazil</i>	Richard Harper, <i>UK</i>
Simone Barbosa, <i>Brazil</i>	Andreas Holzinger, <i>Austria</i>
Patrick Baudisch, <i>Germany</i>	Kasper Hornbaek, <i>Denmark</i>
Regina Bernhaupt, <i>France</i>	Horst Hortner, <i>Austria</i>
Robert Biddle, <i>Canada</i>	Matt Jones, <i>UK</i>
Jeremy Birnholtz, <i>USA</i>	Anirudha Joshi, <i>India</i>
Kellogg Booth, <i>Canada</i>	Hermann Kaindl, <i>Austria</i>
Gaelle Calvary, <i>France</i>	Evangelos Karapanos, <i>Portugal</i>
Pedro Campos, <i>Portugal</i>	Rick Kazman, <i>USA</i>
Torkil Clemmensen, <i>Denmark</i>	Ute Klotz, <i>Switzerland</i>
Nuno Correia, <i>Portugal</i>	Vassilis Kostakos, <i>Portugal</i>
Enrico Costanza, <i>UK</i>	Masaaki Kurosu, <i>Austria</i>
Joelle Coutaz, <i>France</i>	Ed Lank, <i>Canada</i>
José Creissac Campos, <i>Portugal</i>	Marta Larusdottir, <i>Iceland</i>
Mary Czerwinski, <i>USA</i>	Henry Lieberman, <i>USA</i>
Peter Dannenmann, <i>Germany</i>	Panos Markopolous, <i>The Netherlands</i>
Andy Dearden, <i>UK</i>	Christian Muller, <i>Germany</i>
Anke Dittmar, <i>Germany</i>	Miguel Nacenta, <i>Canada</i>
Ellen Do, <i>USA</i>	Laurence Nigay, <i>France</i>
Gavin Doherty, <i>Ireland</i>	Monique Noirhomme, <i>Belgium</i>
Andrew Duchowski, <i>USA</i>	Eamonn O'Neill, <i>UK</i>
Henry Been-Lim Duh, <i>Singapore</i>	Ian Oakley, <i>Portugal</i>
Michael Feary, <i>USA</i>	Oscar Pastor, <i>Spain</i>
Peter Forbrig, <i>Germany</i>	Fabio Paterno, <i>Italy</i>
Nahum Gershon, <i>The Netherlands</i>	Lia Patrício, <i>Portugal</i>
Marianne Graves Petersen, <i>Denmark</i>	Helen Petrie, <i>UK</i>

Nitendra Rajput, *India*

Janet Read, *UK*

Dave Roberts, *UK*

Kari-Jouko Raiha, *Finland*

Miguel Sales Dias, *Portugal*

Jaime Sanchez, *Chile*

Robert St Amant, *USA*

Kamelia Stefanova, *Bulgaria*

James Stewart, *Canada*

Wolfgang Stuerzlinger, *UK*

Jan van den Bergh, *Belgium*

Gerrit van der Veer, *The Netherlands*

Jos van Leeuwen, *Portugal*

Gerhard Weber, *Germany*

Janet Wesson, *South Africa*

Marco Winckler, *France*

Volker Wulf, *Germany*

Associate Chairs - Short Papers

Jose Abdelnour-Nocera, *UK*

Elisabeth André, *Germany*

Mark Apperley, *New Zealand*

Nathalie Aquino, *Spain*

Simone Barbosa, *Brazil*

Alexander Boden, *Germany*

Gaelle Calvary, *France*

Robert Capra, *USA*

Luis Carriço, *Portugal*

Marc Cavazza, *UK*

Teresa Chambel, *Portugal*

Stéphane Conversy, *France*

Nuno Correia, *Portugal*

Tim Davis, *USA*

Antonella de Angeli, *UK*

Andy Dearden, *UK*

Anke Dittmar, *Germany*

Carlos Duarte, *Portugal*

Achim Eber, *Germany*

David Elsweiler, *UK*

Danyel Fisher, *USA*

Peter Forbrig, *Germany*

Tiago Guerreiro, *Portugal*

Jacek Gwizdka, *USA*

Marc Hassenzahl, *Germany*

Anirudha Joshi, *India*

Hermann Kaindl, *Austria*

Ute Klotz, *Switzerland*

Tessa Lau, *USA*

Gitte Lindgaard, *Canada*

Floyd Mueller, *USA*

Lennart Nacke, *Canada*

Yukiko Nakano, *Japan*

Monique Noirhomme, *Belgium*

Lars Oestreicher, *Sweden*

Eamonn O'Neill, *UK*

Dan Orwa, *Kenya*

Tim Paek, *USA*

Ignacio Panach, *Spain*

Fabio Paterno, *Italy*

Lia Patrício, *Portugal*

Nitendra Rajput, *India*

Francisco Rebelo, *Portugal*

Dave Roberts, *UK*

Teresa Romão, *Portugal*

Virpi Roto, *Finland*

Raquel Santos, *Portugal*

Beatriz Sousa Santos, *Portugal*

James Stewart, *Canada*

Sriram Subramanian, *UK*

Feng Tian, *China*

Manas Tungare, *USA*

Gerhard Weber, *Germany*

Astrid Weiss, *Austria*

Marco Winckler, *France*

Chui Yin Wong, *Malaysia*

Reviewers

Al Mahmud Abdullah, *The Netherlands*
 Ana Paula Afonso, *Portugal*
 Jason Alexander, *UK*
 Jan Alexandersson, *Germany*
 Dzmityr Aliakseyeu, *The Netherlands*
 Majed Alshamari, *Saudi Arabia*
 Margarita Anastassova, *France*
 Craig Anslow, *New Zealand*
 Caroline Appert, *France*
 Nathalie Aquino, *Spain*
 Pedro Arezes, *Portugal*
 Ernesto Arroyo, *USA*
 Mark Ashdown, *UK*
 Ching man Au Yeung, *Japan*
 Chris Baber, *UK*
 Paula M. Bach, *USA*
 Nilufar Baghaei, *New Zealand*
 Sebastiano Bagnara, *Italy*
 Gilles Bailly, *Germany*
 Martina Balestra, *USA*
 Emilia Barakova, *The Netherlands*
 Jakob Bardram, *Denmark*
 Shaowen Bardzell, *USA*
 Javier Bargas-Avila, *Switzerland*
 Louise Barkhuus, *Denmark*
 Pippin Barr, *Denmark*
 Barbara Rita Barricelli, *Italy*
 Gil Barros, *Brazil*
 Len Bass, *USA*
 Remi Bastide, *France*
 Rafael Bastos, *Portugal*
 Eric Baumer, *USA*
 Gordon Baxter, *UK*
 Michel Beaudouin-Lafon, *France*
 Nikolaus Bee, *Germany*
 Yacine Bellik, *France*
 Kawtar Benghazi, *Spain*
 Mike Bennett, *USA*
 François Bérard, *France*
 Olav W. Bertelsen, *Denmark*
 Nigel Bevan, *UK*
 Ganesh Bhutkar, *India*
 Matthew Bietz, *USA*
 Mark Billingham, *New Zealand*
 Dorrit Billman, *USA*
 Fernando Birra, *Portugal*
 Mike Blackstock, *Canada*
 Marcus Bloice, *Austria*
 Marco Blumendorf, *Germany*
 Mark Blythe, *UK*
 Cristian Bogdan, *Sweden*
 Morten Bohøj, *Denmark*
 Matthew Bolton, *USA*
 Birgit Bomsdorf, *Germany*
 Rodrigo Bonacin, *Brazil*
 Sebastian Boring, *Canada*
 Aviaja Borup, *Denmark*
 Matt-Mouley Bouamrane, *UK*
 Doug Bowman, *USA*
 Giorgio Brajnik, *Italy*
 Pedro Branco, *Portugal*
 Willem-Paul Brinkman, *The Netherlands*
 Gregor Broll, *Germany*
 Christopher Brooks, *Canada*
 Judith Brown, *Canada*
 Steffen Budweg, *Germany*
 Lucy Buykx, *UK*
 Marina Buzzi, *Italy*
 Daragh Byrne, *Ireland*
 Cristina Cachero, *Spain*
 Jeff Calcaterra, *USA*
 Licia Calvi, *The Netherlands*
 Eduardo Calvillo Gamez, *Mexico*
 Maria-Dolores Cano, *Spain*
 Xiang Cao, *China*
 Cinzia Cappiello, *Italy*
 Robert Capra, *USA*
 Luis Carlos paschoarelli, *Brazil*
 Stefan Carmien, *Spain*
 Maria Beatriz Carmo, *Portugal*
 António Carvalho Brito, *Portugal*
 Luis Castro, *Mexico*
 Daniel Cernea, *Germany*
 Matthew Chalmers, *UK*

Teresa Chambel, *Portugal*
Beenish Chaudry, *USA*
Tao Chen, *China*
Fanny Chevalier, *Canada*
Keith Cheverst, *UK*
Yoram Chisik, *Portugal*
Yu-kwong Chiu, *China*
Georgios Christou, *Cyprus*
Andrea Civan Hartzler, *USA*
Laurence Claeys, *France*
Luis Coelho, *Portugal*
François Coldefy, *France*
Karin Coninx, *Belgium*
Maria Francesca Costabile, *Italy*
Céline Coutrix, *France*
Nadine Couture, *France*
Anna Cox, *UK*
David Coyle, *Ireland*
Leonardo Cunha de Miranda, *Portugal*
Edward Cutrell, *India*
Raimund Dachzelt, *Germany*
José Danado, *Norway*
Tjerk de Greef, *The Netherlands*
Alexander De Luca, *Germany*
Luigi De Russis, *Italy*
Clarisse de Souza, *Brazil*
Alexandre Demeure, *France*
Charlie DeTar, *USA*
Ines Di Loreto, *Italy*
Eduardo Dias, *Portugal*
Paulo Dias, *Portugal*
Claire Diederich, *Belgium*
Andre Doucette, *Canada*
Carlos Duarte, *Portugal*
Emmanuel Dubois, *France*
Cathy Dudek, *Canada*
Andreas Duenser, *New Zealand*
Mark Dunlop, *UK*
Sophie Dupuy-Chessa, *France*
Matthew Easterday, *USA*
Achim Ebert, *Germany*
Florian Echtler, *USA*
Amnon Eden, *UK*
Serge Egelman, *USA*
Linda Elliott, *USA*
Niklas Elmqvist, *USA*
Alex Endert, *USA*
Dominik Ertl, *Austria*
Parisa Eslambolchilar, *UK*
Augusto Esteves, *Portugal*
Pedro Faria Lopes, *Portugal*
Robert Farrell, *USA*
Ian Fasel, *USA*
Ava Fatah gen. Schieck, *UK*
Jean-Daniel Fekete, *France*
Xavier Ferre, *Spain*
Mirko Fetter, *Germany*
Sebastian Feuerstack, *Brazil*
Nelson Figueiredo de Pinho, *Portugal*
George Fitzmaurice, *Canada*
Joan Fons, *Spain*
Manuel J. Fonseca, *Portugal*
Alain Forget, *Canada*
Florian Förster, *Austria*
Derek Foster, *UK*
Marcus Foth, *Australia*
Teresa Franqueira, *Portugal*
Mike Fraser, *UK*
Christopher Frauenberger, *UK*
André Freire, *UK*
Carla Freitas, *Brazil*
David Frohlich, *UK*
Dominic Furniss, *UK*
Luigi Gallo, *Italy*
Teresa Galvão, *Portugal*
Nestor Garay-Vitoria, *Spain*
Roberto García, *Spain*
Anant Bhaskar Garg, *India*
Vaibhav Garg, *USA*
Jose Luis Garrido, *Spain*
Nahum Gershon, *Canada*
Florian Geyer, *Germany*
Werner Geyer, *USA*
Giuseppe Ghiani, *Italy*
Andy Gimblett, *UK*
Patrick Girard, *France*
Sylvie Girard, *UK*
Leonardo Giusti, *Italy*
Guilherme Gomes, *Portugal*
Daniel Gonçalves, *Portugal*

- José Luis González Sánchez, *Spain*
Phil Gosset, *UK*
Nitesh Goyal, *USA*
Toni Granollers, *Spain*
Anders Green, *Sweden*
Collin Green, *USA*
Saul Greenberg, *Canada*
Olivier Grisvard, *France*
Tiago Guerreiro, *Portugal*
Sean Gustafson, *Germany*
Mieke Haesen, *Belgium*
Jonna Häkkinen, *Finland*
Martin Halvey, *UK*
Judy Hammond, *Australia*
Mark Hancock, *Canada*
Morten Borup Harning, *Denmark*
John Harris, *Canada*
Kirstie Hawkey, *Canada*
Elaine Hayashi, *Brazil*
Brent Hecht, *USA*
Steffen Hedegaard, *Denmark*
Mathias Heilig, *Germany*
Ruediger Heimgaertner, *Germany*
Ingi Helgason, *UK*
Sarah Henderson, *New Zealand*
Bart Hengeveld, *The Netherlands*
Wilko Heuten, *Germany*
Michael Hildebrandt, *Norway*
Christina Hochleitner, *Austria*
Eve Hoggan, *Finland*
Paul Holleis, *Germany*
Clemens Holzmann, *Austria*
Jettie Hoonhout, *The Netherlands*
Michael Horn, *USA*
Eva Hornecker, *Germany*
Heiko Hornung, *Brazil*
Horst Hörtnner, *Austria*
Juan Pablo Hourcade, *USA*
Aaron Houssian, *The Netherlands*
Andrew Howes, *UK*
Dalibor Hrg, *Germany*
Ko-Hsun Huang, *Portugal*
Jina Huh, *USA*
Tim Hussein, *Germany*
Dugald Hutchings, *USA*
Junko Ichino, *Japan*
Netta Iivari, *Finland*
Emilio Insfran, *Spain*
Samuel Inverso, *Australia*
Shamsi Iqbal, *USA*
Petra Isenberg, *France*
Howell Istance, *UK*
Linda Jackson, *USA*
Robert Jacob, *USA*
Mikkel Jakobsen, *Denmark*
Jacek Jankowski, *USA*
Hans-Christian Jetter, *Germany*
Sune Alstrup Johansen, *Denmark*
Jeff Johnson, *USA*
Simon Jones, *UK*
Martino Jose Mario, *Brazil*
Rui José, *Portugal*
Marko Jurmu, *Finland*
Don Kalar, *USA*
Vaiva Kalnikaite, *UK*
Martin Kaltenbrunner, *Austria*
Matthew Kam, *USA*
Mayur Karnik, *Portugal*
Hannu Karvonen, *Finland*
Sebastian Kassner, *Germany*
Dinesh Katre, *India*
Sevan Kavaldjian, *Austria*
Konstantinos Kazakos, *Australia*
Pramod Khambete, *India*
Vassilis-Javed Khan, *The Netherlands*
Hyungsin Kim, *USA*
Jayne Klenner-Moore, *USA*
Christian Kray, *UK*
Per Ola Kristensson, *UK*
Hannu Kukka, *Finland*
Andrew Kun, *USA*
H. Chad Lane, *USA*
Yann Laurillau, *France*
Effie Law, *Switzerland*
Marco Lazzari, *Italy*
Karin Leichtenstern, *Germany*
Juha Leino, *Finland*
Barbara Leporini, *Italy*
Sophie Lepreux, *France*
Olivier Lequenne, *France*

Chunyuan Liao, *USA*
Conor Linehan, *UK*
Agnes Lisowska Masson, *China*
Zhengjie Liu, *China*
Sara Ljungblad, *Sweden*
Claire Lobet, *Belgium*
Steffen Lohmann, *Spain*
Fernando Lopez-Colino, *Spain*
Anja Lorenz, *Germany*
Stephanie Ludi, *USA*
Bernd Ludwig, *Germany*
Andreas Luedtke, *Germany*
Jo Lumsden, *UK*
Kris Luyten, *Belgium*
Kent Lyons, *Canada*
Allan MacLean, *UK*
Joaquim Madeira, *Portugal*
Rui Madeira, *Portugal*
Angela Mahr, *Germany*
Stephann Makri, *UK*
Sylvain Malacria, *France*
Benjamin Mangold, *Germany*
Javier Marco, *Spain*
Gary Marsden, *South Africa*
Mark Marshall, *UK*
Hannah Marston, *Canada*
Jean-Bernard Martens,
The Netherlands
Lynne Martin, *USA*
Diego Martínez, *Spain*
Célia Martinie, *France*
Masood Massodian, *New Zealand*
Sara Mastro, *USA*
Maristella Matera, *Italy*
Akhil Mathur, *Canada*
Eva Mayr, *Austria*
Davide Mazza, *Italy*
emanuela mazzone, *UK*
Gregor McEwan, *Canada*
Kevin McGee, *Singapore*
Marilyn McGee-Lennon, *UK*
Indrani Medhi, *India*
Gerrit Meixner, *Germany*
Guy Melancon, *France*
Eduarda Mendes Rodrigues, *Portugal*
Helena Mentis, *UK*
Tim Merritt, *Singapore*
Mei Miao, *Germany*
Alex Mitchell, *Singapore*
Robb Mitchell, *Denmark*
Jose Pascual Molina Masso, *Spain*
Francisco Montero, *Spain*
Meredith Morris, *USA*
Ann Morrison, *Denmark*
Christiane Moser, *Austria*
Omar Mubin, *The Netherlands*
Florian 'Floyd' Mueller, *USA*
Christian Mueller-Tomfelde, *Australia*
Michael Muller, *USA*
Maurice Mulvenna, *UK*
Dianne Murray, *UK*
Lennart Nacke, *Canada*
Peyman Nasirifard, *USA*
David Navarre, *France*
Ather Nawaz, *Denmark*
Luciana Nedel, *Brazil*
Vania Neris, *Brazil*
Colette Nicolle, *UK*
Femke Nijboer, *The Netherlands*
Valentina Nisi, *Portugal*
Leonel Nobrega, *Portugal*
Sylvie Noel, *Canada*
Manuel Noguera, *Spain*
Marianna Obrist, *Austria*
Johanna Renny Octavia, *Belgium*
Amy Ogan, *USA*
Michael O'Grady, *Ireland*
Kenton O'Hara, *UK*
Timo Ojala, *Finland*
Eugenio Oliveira, *Portugal*
Veronica Orvalho, *Portugal*
Nuno Otero, *Portugal*
Benoit Otjacques, *Luxembourg*
Ana Paiva, *Portugal*
Yue Pan, *USA*
Jose Ignacio Panach Navarrete, *Spain*
Alex Pang, *UK*
Nadia Pantidi, *UK*
Luca Paolino, *Italy*
Eleftherios Papachristos, *Greece*

Narcis Pares, *USA*
Andrew Patrick, *Canada*
Celeste Lyn Paul, *USA*
Sharoda Paul, *USA*
Andriy Pavlovych, *Canada*
Greg Phillips, *Canada*
Lara Piccolo, *Brazil*
Martin Pielot, *Germany*
Emmanuel Pietriga, *France*
franck poirier, *France*
Benjamin Poppinga, *Germany*
Christopher Power, *UK*
Raquel Prates, *Brazil*
John Precious, *UK*
Costin Pribeanu, *Romania*
Andreas Pusch, *France*
Alexandra Queirós, *Portugal*
Ismo Rakkolainen, *Finland*
Dave Randall, *UK*
Alberto Raposo, *Brazil*
Stuart Reeves, *UK*
Patrick Reignier, *France*
René Reiners, *Germany*
Malte Ressin, *UK*
Bernardo Reynolds, *Portugal*
Andy Ridge, *UK*
Xavier Righetti, *Switzerland*
Pierre Robillard, *Canada*
Simon Robinson, *UK*
Carsten Röcker, *Germany*
Yvonne Rogers, *UK*
Markus Rohde, *Germany*
Teresa Romão, *Portugal*
Virpi Roto, *Finland*
Anne Roudaut, *Germany*
jose rouillard, *France*
Mark Rouncefield, *UK*
Nicolas Roussel, *France*
Jaime Ruiz, *Canada*
Pascal Salembier, *France*
Antti Salovaara, *Finland*
Nithya Sambasivan, *USA*
Krystian Samp, *Ireland*
Paulo Sampaio, *Portugal*
Vagner Santana, *Italy*
Carmen Santoro, *Italy*
José Santos, *Portugal*
Teresa Sarmiento, *Portugal*
Cheryl Savery, *Canada*
Dominique Scapin, *France*
Thomas Schlegel, *Germany*
Kevin Schneider, *Canada*
Johannes Schöning, *Germany*
Eric Schweikardt, *USA*
Gig Searle, *Austria*
Thomas Seifried, *Austria*
Marc Seissler, *Germany*
Malu Seixas, *Brazil*
Ted Selker, *USA*
Abi Sellen, *UK*
Dev Sen, *Canada*
Andrew Seniuk, *Canada*
Aaditeshwar Seth, *India*
Leslie Setlock, *USA*
Ehud Sharlin, *Canada*
Aditi Sharma, *South Africa*
Huihui Shi, *Germany*
Aubrey Shick, *USA*
Garth Shoemaker, *Canada*
Bruno Silva, *Brazil*
Frutuoso Silva, *Portugal*
Hugo Silva, *Portugal*
Klaus-Martin Simonic, *Austria*
Mikael B. Skov, *Denmark*
Roger Slack, *UK*
David Smith, *Canada*
Dustin Smith, *USA*
Thomas Smyth, *Canada*
William Soukoreff, *Canada*
Kenia Sousa, *Belgium*
Jan Stage, *Denmark*
Danae Stanton Fraser, *UK*
Gunnar Stevens, *Germany*
Erik Stolterman, *USA*
Markus Stolze, *Switzerland*
Steven Strachan, *USA*
Simone Stumpf, *UK*
Sriram Subramanian, *UK*
Ja-Young Sung, *USA*
Alistair Sutcliffe, *UK*

David Swallow, *UK*
Colin Swindells, *Canada*
Gerd Szwillus, *Germany*
Susanne Tak, *New Zealand*
Anthony Tang, *USA*
Charlotte Tang, *Canada*
Michael Tangermann, *Germany*
Franck Tarpin-Bernard, *France*
Alex Taylor, *UK*
Stephanie Teasley, *USA*
António Teixeira, *Portugal*
Michael Terry, *Canada*
VinhTuan Thai, *Ireland*
Harold Thimbleby, *UK*
Martin Tomitsch, *Australia*
Daniela Trevisan, *Brazil*
Sylvia Truman, *UK*
Manfred Tscheligi, *Austria*
Nikolaos Tselios, *Greece*
Simon Tucker, *UK*
Markku Turunen, *Finland*
Brygg Ullmer, *USA*
Leon Urbas, *Germany*
Teija Vainio, *Finland*
Leonel Valbom, *Portugal*
Egon L. van den Broek, *Austria*
Thea van der Geest, *The Netherlands*
Ielka van der Sluis, *Ireland*
Erik van der Spek, *The Netherlands*
Jean Vanderdonckt, *Belgium*
Radu-Daniel Vatavu, *Romania*
Manuel Veit, *France*
Jayant Venkatanathan, *Portugal*
Arnold P.O.S. Vermeeren,
The Netherlands
Bart Vermeersch, *Belgium*
Jo Vermeulen, *Belgium*
Frédéric Vernier, *France*
Roel Vertegaal, *Canada*
Markel Vigo, *UK*
Nadine Vigouroux, *France*
Thomas Visser, *The Netherlands*
Stephen Voida, *USA*
Ivan Volosyak, *Germany*
Jade Wang, *USA*
Qing Wang, *China*
Leon Watts, *UK*
Astrid Weiss, *Austria*
Peter Wild, *UK*
Graham Wilson, *UK*
Max Wilson, *UK*
Heike Winschiers-Theophilus, *Namibia*
Jacob Wobbrock, *USA*
Peter Wolkerstorfer, *Austria*
Chui Yin Wong, *Malaysia*
Michael Wright, *UK*
Min Wu, *USA*
Peta Wyeth, *Australia*
Alvin W. Yeo, *Malaysia*
James Young, *Canada*
Ray Yun, *USA*
Loutfouz Zaman, *Canada*
Panayiotis Zaphiris, *Cyprus*
Martina Ziefle, *Germany*
Juergen Ziegler, *Germany*
Gottfried Zimmermann, *Germany*
Martin Zimmermann, *Germany*

Sponsors

Gold

Microsoft®
Research

Silver



FCT

Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Bronze



YDREAMS™

Supporters



Organization

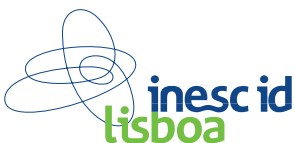


Table of Contents – Part I

Keynote Speakers

Natural User Interfaces	1
<i>Antônio Câmara</i>	
The Future of Distributed Groups and Their Use of Social Media	2
<i>Mary Czerwinski</i>	
Opportunities for Proxemic Interactions in Ubicomp (Keynote)	3
<i>Saul Greenberg</i>	

Long and Short Papers

Accessibility I

Voice Games: Investigation Into the Use of Non-speech Voice Input for Making Computer Games More Accessible	11
<i>Susumu Harada, Jacob O. Wobbrock, and James A. Landay</i>	
GraVVITAS: Generic Multi-touch Presentation of Accessible Graphics	30
<i>Cagatay Goncu and Kim Marriott</i>	
Designing a Playful Communication Support Tool for Persons with Aphasia	49
<i>Abdullah Al Mahmud, Idowu I.B.I. Ayoola, and Jean-Bernard Martens</i>	
How to Make Numerical Information Accessible: Experimental Identification of Simplification Strategies	57
<i>Susana Bautista, Raquel Hervás, Pablo Gervás, Richard Power, and Sandra Williams</i>	

Accessibility II

Blind People and Mobile Keypads: Accounting for Individual Differences	65
<i>Tiago Guerreiro, João Oliveira, João Benedito, Hugo Nicolau, Joaquim Jorge, and Daniel Gonçalves</i>	
Elderly User Evaluation of Mobile Touchscreen Interactions	83
<i>Masatomo Kobayashi, Atsushi Hiyama, Takahiro Miura, Chieko Asakawa, Michitaka Hirose, and Tohru Ifukube</i>	

BrailleType: Unleashing Braille over Touch Screen Mobile Phones	100
<i>João Oliveira, Tiago Guerreiro, Hugo Nicolau, Joaquim Jorge, and Daniel Gonçalves</i>	
Potential Pricing Discrimination Due to Inaccessible Web Sites	108
<i>Jonathan Lazar, Brian Wentz, Matthew Bogdan, Edrick Clowney, Matthew Davis, Joseph Guiffo, Danial Gunnarsson, Dustin Hanks, John Harris, Behnjay Holt, Mark Kitchin, Mark Motayne, Roslin Nzokou, Leela Sedaghat, and Kathryn Stern</i>	
Affective HCI	
Measuring Immersion and Affect in a Brain-Computer Interface Game	115
<i>Gido Hakvoort, Hayrettin Gürkök, Danny Plass-Oude Bos, Michel Obbink, and Mannes Poel</i>	
Understanding Goal Setting Behavior in the Context of Energy Consumption Reduction	129
<i>Michelle Scott, Mary Barreto, Filipe Quintal, and Ian Oakley</i>	
Designing a Context-Aware Architecture for Emotionally Engaging Mobile Storytelling	144
<i>Fabio Pittarello</i>	
Towards Emotional Interaction: Using Movies to Automatically Learn Users' Emotional States	152
<i>Eva Oliveira, Mitchel Benovoy, Nuno Ribeiro, and Teresa Chambel</i>	
Computer-Mediated Communication	
Motion and Attention in a Kinetic Videoconferencing Proxy	162
<i>David Sirkin, Gina Venolia, John Tang, George Robertson, Taemie Kim, Kori Inkpen, Mara Sedlins, Bongshin Lee, and Mike Sinclair</i>	
Making Sense of Communication Associated with Artifacts during Early Design Activity	181
<i>Moushumi Sharmin and Brian P. Bailey</i>	
Children's Interactions in an Asynchronous Video Mediated Communication Environment	199
<i>Michail N. Giannakos, Konstantinos Chorianopoulos, Paul Johns, Kori Inkpen, and Honglu Du</i>	
Effects of Automated Transcription Delay on Non-native Speakers' Comprehension in Real-Time Computer-Mediated Communication	207
<i>Lin Yao, Ying-xin Pan, and Dan-ning Jiang</i>	

Computer-Supported Cooperative Work I

- Redundancy and Collaboration in Wikibooks 215
*Ilaria Liccardi, Olivier Chapuis, Ching-Man Au Yeung, and
 Wendy Mackay*
- Towards Interoperability in Municipal Government: A Study of
 Information Sharing Practices 233
*Stacy F. Hobson, Rangachari Anand, Jeaha Yang, and
 Juhnyoung Lee*
- An Integrated Communication and Collaboration Platform for
 Distributed Scientific Workgroups 248
Christian Müller-Tomfelde, Jane Li, and Alex Hyatt

Computer-Supported Cooperative Work II

- IdeaTracker: An Interactive Visualization Supporting Collaboration
 and Consensus Building in Online Interface Design Discussions 259
*Roshanak Zilouchian Moghaddam, Brian P. Bailey, and
 Christina Poon*
- What You See Is What You (Can) Get? Designing for Process
 Transparency in Financial Advisory Encounters 277
Philipp Nussbaumer and Inu Matter
- A Framework for Supporting Joint Interpersonal Attention in
 Distributed Groups 295
*Jeremy Birnholtz, Johnathon Schultz, Matthew Lepage, and
 Carl Gutwin*

Evaluation I

- Do Teams Achieve Usability Goals? Evaluating Goal Achievement with
 Usability Goals Setting Tool 313
Anirudha Joshi and N.L. Sarda
- Supporting Window Switching with Spatially Consistent Thumbnail
 Zones: Design and Evaluation 331
Susanne Tak, Joey Scarr, Carl Gutwin, and Andy Cockburn
- Evaluating Commonsense Knowledge with a Computer Game 348
Juan F. Mancilla-Caceres and Eyal Amir
- Remote Usability Testing Using Eyetracking 356
Piotr Chynal and Jerzy M. Szymański

Evaluation II

A Means-End Analysis of Consumers' Perceptions of Virtual World Affordances for E-commerce	362
<i>Minh Quang Tran, Shailey Minocha, Dave Roberts, Angus Laing, and Darren Langdrige</i>	
Improving Users' Consistency When Recalling Location Sharing Preferences	380
<i>Jayant Venkatanathan, Denzil Ferreira, Michael Benisch, Jialiu Lin, Evangelos Karapanos, Vassilis Kostakos, Norman Sadeh, and Eran Toch</i>	
Navigation Time Variability: Measuring Menu Navigation Errors	388
<i>Krystian Samp and Stefan Decker</i>	
Challenges in Designing Inter-usable Systems	396
<i>Ville Antila and Alfred Lui</i>	
Directed Cultural Probes: Detecting Barriers in the Usage of Public Transportation	404
<i>Susanne Schmehl, Stephanie Deutsch, Johann Schrammel, Lucas Paletta, and Manfred Tscheligi</i>	

Finding and Retrieving

Image Retrieval with Semantic Sketches	412
<i>David Engel, Christian Herdtweck, Björn Browatzki, and Cristóbal Curio</i>	
Mixer: Mixed-Initiative Data Retrieval and Integration by Example	426
<i>Steven Gardiner, Anthony Tomasic, John Zimmerman, Rafae Aziz, and Kathryn Rivard</i>	
Speaking to See: A Feasibility Study of Voice-Assisted Visual Search . . .	444
<i>Victor Kaptelinin and Herje Wåhlen</i>	

Fun / Aesthetic Design I

Analysing the Playground: Sensitizing Concepts to Inform Systems That Promote Playful Interaction	452
<i>Stefan Rennick Egglestone, Brendan Walker, Joe Marshall, Steve Benford, and Derek McAuley</i>	

Comparative Feedback in the Street: Exposing Residential Energy Consumption on House Façades	470
<i>Andrew Vande Moere, Martin Tomitsch, Monika Hoinkis, Elmar Trefz, Silje Johansen, and Allison Jones</i>	
Are First Impressions about Websites Only Related to Visual Appeal?	489
<i>Eleftherios Papachristos and Nikolaos Avouris</i>	
You Can Wear It, But Do They Want to Share It or Stare at It?	497
<i>Arto Puikkonen, Anu Lehtiö, and Antti Virolainen</i>	

Fun / Aesthetic Design II

Design and Evaluation of Interaction Technology for Medical Team Meetings	505
<i>Alex Olwal, Oscar Frykholm, Kristina Groth, and Jonas Moll</i>	
How Technology Influences the Therapeutic Process: A Comparative Field Evaluation of Augmented Reality and In Vivo Exposure Therapy for Phobia of Small Animals	523
<i>Maja Wrzesien, Jean-Marie Burkhardt, Mariano Alcañiz, and Cristina Botella</i>	
You've Covered: Designing for In-Shift Handoffs in Medical Practice	541
<i>Yunan Chen</i>	

Gestures

A Taxonomy of Microinteractions: Defining Microgestures Based on Ergonomic and Scenario-Dependent Requirements	559
<i>Katrin Wolf, Anja Naumann, Michael Rohs, and Jörg Müller</i>	
Unifying Events from Multiple Devices for Interpreting User Intentions through Natural Gestures	576
<i>Pablo Llinás, Manuel García-Herranz, Pablo A. Haya, and Germán Montoro</i>	
SimpleFlow: Enhancing Gestural Interaction with Gesture Prediction, Abbreviation and Autocompletion	591
<i>Mike Bennett, Kevin McCarthy, Sile O'Modhrain, and Barry Smyth</i>	

HCI in the Classroom

The Perception of Sound and Its Influence in the Classroom	609
<i>Sofia Reis and Nuno Correia</i>	

Encouraging Initiative in the Classroom with Anonymous Feedback	627
<i>Tony Bergstrom, Andrew Harris, and Karrie Karahalios</i>	
U-Note: Capture the Class and Access It Everywhere	643
<i>Sylvain Malacria, Thomas Pietrzak, Aurélien Tabard, and Éric Lecolinet</i>	

Erratum

Design and Evaluation of Interaction Technology for Medical Team Meetings	E1
<i>Alex Olwal, Oscar Frykholm, Kristina Groth, and Jonas Moll</i>	
Author Index	661

Table of Contents – Part II

Long and Short Papers

Health I

Finding the Right Way for Interrupting People Improving Their Sitting Posture	1
<i>Michael Haller, Christoph Richter, Peter Brandl, Sabine Gross, Gerold Schossleitner, Andreas Schrempf, Hideaki Nii, Maki Sugimoto, and Masahiko Inami</i>	
Exploring Haptic Feedback in Exergames	18
<i>Tadeusz Stach and T.C. Nicholas Graham</i>	
Identifying Barriers to Effective User Interaction with Rehabilitation Tools in the Home	36
<i>Stephen Uzor, Lynne Baillie, Dawn Skelton, and Fiona Fairlie</i>	
Clinical Validation of a Virtual Environment Test for Safe Street Crossing in the Assessment of Acquired Brain Injury Patients with and without Neglect	44
<i>Patricia Mesa-Gresa, Jose A. Lozano, Roberto Llórens, Mariano Alcañiz, María Dolores Navarro, and Enrique Noé</i>	

Health II

Smart Homes or Smart Occupants? Supporting Aware Living in the Home	52
<i>Lyn Bartram, Johnny Rodgers, and Rob Woodbury</i>	
Input Devices in Mental Health Applications: Steering Performance in a Virtual Reality Paths with WiiMote	65
<i>Maja Wrzesien, María José Rupérez, and Mariano Alcañiz</i>	
Ácted RealityŠ in Electronic Patient Record Research: A Bridge between Laboratory and Ethnographic Studies	73
<i>Lesley Axelrod, Geraldine Fitzpatrick, Flis Henwood, Liz Thackray, Becky Simpson, Amanda Nicholson, Helen Smith, Greta Rait, and Jackie Cassell</i>	
Exercise Support System for Elderly: Multi-sensor Physiological State Detection and Usability Testing	81
<i>Jan Macek and Jan Kleindienst</i>	

Human Factors I

Estimating the Perceived Difficulty of Pen Gestures	89
<i>Radu-Daniel Vatavu, Daniel Vogel, Géry Casiez, and Laurent Grisoni</i>	
On the Limits of the Human Motor Control Precision: The Search for a Device's Human Resolution	107
<i>François Bérard, Guangyu Wang, and Jeremy R. Cooperstock</i>	
Three around a Table: The Facilitator Role in a Co-located Interface for Social Competence Training of Children with Autism Spectrum Disorder	123
<i>Massimo Zancanaro, Leonardo Giusti, Eynat Gal, and Patrice T. Weiss</i>	

Human Factors II

Moving Target Selection in 2D Graphical User Interfaces	141
<i>Abir Al Hajri, Sidney Fels, Gregor Miller, and Michael Ilich</i>	
Navigational User Interface Elements on the Left Side: Intuition of Designers or Experimental Evidence?	162
<i>Andreas Holzinger, Reinhold Scherer, and Martina Ziefle</i>	
Pupillary Response Based Cognitive Workload Measurement under Luminance Changes	178
<i>Jie Xu, Yang Wang, Fang Chen, and Eric Choi</i>	
Study on the Usability of a Haptic Menu for 3D Interaction	186
<i>Giandomenico Caruso, Elia Gatti, and Monica Bordegoni</i>	

Interacting in Public Spaces

Balancing Act: Enabling Public Engagement with Sustainability Issues through a Multi-touch Tabletop Collaborative Game	194
<i>Alissa N. Antle, Joshua Tanenbaum, Allen Bevans, Katie Seaborn, and Sijie Wang</i>	
Understanding the Dynamics of Engaging Interaction in Public Spaces	212
<i>Peter Dalsgaard, Christian Dindler, and Kim Halskov</i>	
Transferring Human-Human Interaction Studies to HRI Scenarios in Public Space	230
<i>Astrid Weiss, Nicole Mirnig, Roland Buchner, Florian Förster, and Manfred Tscheligi</i>	

Interacting with Displays

Comparing Free Hand Menu Techniques for Distant Displays Using Linear, Marking and Finger-Count Menus	248
<i>Gilles Bailly, Robert Walter, Jörg Müller, Tongyan Ning, and Eric Lecolinet</i>	
Design and Evaluation of an Ambient Display to Support Time Management during Meetings	263
<i>Valentina Occhialini, Harm van Essen, and Berry Eggen</i>	
Does Panel Type Matter for LCD Monitors? A Study Examining the Effects of S-IPS, S-PVA, and TN Panels in Video Gaming and Movie Viewing	281
<i>Ki Joon Kim and S. Shyam Sundar</i>	
ModControl – Mobile Phones as a Versatile Interaction Device for Large Screen Applications	289
<i>Matthias Deller and Achim Ebert</i>	

Interaction Design for Developing Regions

A New Visualization Approach to Re-Contextualize Indigenous Knowledge in Rural Africa	297
<i>Kasper Rodil, Heike Winschiers-Theophilus, Nicola J. Bidwell, Søren Eskildsen, Matthias Rehm, and Gereon Koch Kapuire</i>	
Design Opportunities for Supporting Treatment of People Living with HIV / AIDS in India	315
<i>Anirudha Joshi, Mandar Rane, Debjani Roy, Shweta Sali, Neha Bharshankar, N. Kumarasamy, Sanjay Pujari, Davidson Solomon, H. Diamond Sharma, D.G. Saple, Romain Rutten, Aakash Ganju, and Joris Van Dam</i>	
In Class Adoption of Multimedia Mobile Phones by Gender - Results from a Field Study	333
<i>Elba del Carmen Valderrama-Bahamondez, Jarmo Kauko, Jonna Häkkinä, and Albrecht Schmidt</i>	

Interface Design

Scenarchitectures: The Use of Domain-Specific Architectures to Bridge Design and Implementation	341
<i>Nicholas Graham, Emmanuel Dubois, Christophe Bortolaso, and Christopher Wolfe</i>	
Pattern Tool Support to Guide Interface Design	359
<i>Russell Beale and Behzad Bordbar</i>	

Meerkat and Tuba: Design Alternatives for Randomness, Surprise and Serendipity in Reminiscing 376
John Helmes, Kenton O’Hara, Nicolas Vilar, and Alex Taylor

International and Cultural Aspects of HCI

Culture and Facial Expressions: A Case Study with a Speech Interface 392
Beant Dhillon, Rafal Kocielnik, Ioannis Politis, Marc Swerts, and Dalila Szostak

Equality = Inequality: Probing Equality-Centric Design and Development Methodologies 405
Rilla Khaled

e-Rural: A Framework to Generate Hyperdocuments for Milk Producers with Different Levels of Literacy to Promote Better Quality Milking 422
Vanessa Maia Aguiar de Magalhaes, Junia Coutinho Anacleto, André Bueno, Marcos Alexandre Rose Silva, Sidney Fels, and Fernando Cesar Balbino

Designing Interactive Storytelling: A Virtual Environment for Personal Experience Narratives 430
Ilda Ladeira, Gary Marsden, and Lesley Green

Interruptions and Attention

Choosing Your Moment: Interruptions in Multimedia Annotation 438
Christopher P. Bowers, Will Byrne, Benjamin R. Cowan, Chris Creed, Robert J. Hendley, and Russell Beale

Attention and Intention Goals Can Mediate Disruption in Human-Computer Interaction 454
Ernesto Arroyo and Ted Selker

Again?!! The Emotional Experience of Social Notification Interruptions 471
Celeste Lyn Paul, Anita Komlodi, and Wayne Lutters

Do Not Disturb: Physical Interfaces for Parallel Peripheral Interactions 479
Fernando Olivera, Manuel García-Herranz, Pablo A. Haya, and Pablo Llinás

Mobile Interfaces

Information to Go: Exploring In-Situ Information Pick-Up “In the Wild”	487
<i>Hannu Kukka, Fabio Kruger, Vassilis Kostakos, Timo Ojala, and Marko Jurmu</i>	
IntelliTilt: An Enhanced Tilt Interaction Technique for Mobile Map-Based Applications	505
<i>Bradley van Tonder and Janet Wesson</i>	
Tensions in Developing a Secure Collective Information Practice - The Case of Agile Ridesharing	524
<i>Kenneth Radke, Margot Brereton, Seyed Mirisae, Sunil Ghelawat, Colin Boyd, and Juan Gonzalez Nieto</i>	
Choose Popovers over Buttons for iPad Questionnaires	533
<i>Kevin Gaunt, Felix M. Schmitz, and Markus Stolze</i>	

Multi-Modal Interfaces

Developing and Evaluating a Non-visual Memory Game.....	541
<i>Ravi Kuber, Matthew Tretter, and Emma Murphy</i>	
Playing with Tactile Feedback Latency in Touchscreen Interaction: Two Approaches	554
<i>Topi Kaaresoja, Eve Hoggan, and Emilia Anttila</i>	
The Role of Modality in Notification Performance	572
<i>David Warnock, Marilyn McGee-Lennon, and Stephen Brewster</i>	

Multi-User Interaction / Cooperation

Co-located Collaborative Sensemaking on a Large High-Resolution Display with Multiple Input Devices	589
<i>Katherine Vogt, Lauren Bradel, Christopher Andrews, Chris North, Alex Endert, and Duke Hutchings</i>	
Exploring How Tangible Tools Enable Collaboration in a Multi-touch Tabletop Game	605
<i>Tess Speelpenning, Alissa N. Antle, Tanja Doering, and Elise van den Hoven</i>	
Hidden Details of Negotiation: The Mechanics of Reality-Based Collaboration in Information Seeking	622
<i>Mathias Heilig, Stephan Huber, Jens Gerken, Mischa Demarmels, Katrin Allmendinger, and Harald Reiterer</i>	

Navigation and Wayfinding

A Tactile Compass for Eyes-Free Pedestrian Navigation	640
<i>Martin Pielot, Benjamin Poppinga, Wilko Heuten, and Susanne Boll</i>	
Are We There Yet? A Probing Study to Inform Design for the Rear Seat of Family Cars	657
<i>David Wilfinger, Alexander Meschtscherjakov, Martin Murer, Sebastian Osswald, and Manfred Tscheligi</i>	
Don't Look at Me, I'm Talking to You: Investigating Input and Output Modalities for In-Vehicle Systems	675
<i>Lars Holm Christiansen, Nikolaj Yde Frederiksen, Brit Susan Jensen, Alex Ranch, Mikael B. Skov, and Nissanthen Thiruravichandran</i>	
Author Index	693

Table of Contents – Part III

Long and Short Papers

Novel User Interfaces and Interaction Techniques I

A Framework to Develop VR Interaction Techniques Based on OpenInterface and AFreeCA	1
<i>Diego Martínez, J-Y. Lionel Lawson, José P. Molina, Arturo S. García, Pascual González, Jean Vanderdonckt, and Benoit Macq</i>	
Exploring Interaction Strategies in the Context of Sleep	19
<i>Dzmitry Aliakseyeu, Jia Du, Elly Zwartkruis-Pelgrim, and Sriram Subramanian</i>	
FeetUp: A Playful Accessory to Practice Social Skills through Free-Play Experiences	37
<i>Andrea Rosales, Ernesto Arroyo, and Josep Blat</i>	
Designing <i>Snakey</i> : A Tangible User Interface Supporting Well Path Planning	45
<i>John Harris, James Young, Nicole Sultanum, Paul Lapidés, Ehud Sharlin, and Mario Costa Sousa</i>	

Novel User Interfaces and Interaction Techniques II

OP: A Novel Programming Model for Integrated Design and Prototyping of Mixed Objects	54
<i>Céline Coutrix and Laurence Nigay</i>	
A Personal Approach: The <i>Persona</i> Technique in a Companion's Design Lifecycle	73
<i>Joana Campos and Ana Paiva</i>	
Emotive Expression through the Movement of Interactive Robotic Vehicles	91
<i>Eric Kryski and Ehud Sharlin</i>	

Paper 2.0

Evaluation of an Integrated Paper and Digital Document Management System	100
<i>Matthew Jervis and Masood Masoodian</i>	

BendFlip: Examining Input Techniques for Electronic Book Readers
with Flexible Form Factors 117
Doug Wightman, Tim Ginn, and Roel Vertegaal

Who’s That Girl? Handheld Augmented Reality for Printed Photo
Books 134
Niels Henze and Susanne Boll

Recommender Systems

Looking for “Good” Recommendations: A Comparative Evaluation of
Recommender Systems 152
*Paolo Cremonesi, Franca Garzotto, Sara Negro,
Alessandro Vittorio Papadopoulos, and Roberto Turrin*

All the News That’s Fit to Read: Finding and Recommending News
Online 169
Juha Leino, Kari-Jouko R  ih  , and Sanna Finnberg

Helping Users Sort Faster with Adaptive Machine Learning
Recommendations 187
Steven M. Drucker, Danyel Fisher, and Sumit Basu

Social Media and Privacy

Sharing Ephemeral Information in Online Social Networks: Privacy
Perceptions and Behaviours 204
*Bernardo Reynolds, Jayant Venkatanathan, Jorge Gon  alves, and
Vassilis Kostakos*

An Investigation into Facebook Friend Grouping 216
*Patrick Gage Kelley, Robin Brewer, Yael Mayer,
Lorrie Faith Cranor, and Norman Sadeh*

Privacy Concern and Trust in Using Social Network Sites:
A Comparison between French and Chinese Users 234
Li Chen and Ho Keung Tsoi

Privacy Concerns in Enterprise Social Travel: Attitudes and Actions 242
*Netta Aizenbud-Reshef, Artem Barger, Yael Dubinsky, Ido Guy, and
Shiri Kremer-Davidson*

Social Networks

Online Games and Family Ties: Influences of Social Networking Game
on Family Relationship 250
Jing Wen, Yong Ming Kow, and Yunan Chen

The Influence of Customer Familiarity and Personal Innovativeness toward Information Technologies on the Sense of Virtual Community and Participation	265
<i>Manuel J. Sánchez-Franco, José Antonio Carballar-Falcón, Francisco J. Martínez-López, and Juan Carlos Gázquez-Abad</i>	

Characterizing Interactions among Members of Deaf Communities in Orkut	280
<i>Glúvia A.R. Barbosa, Ismael S. Silva, Glauber Gonçalves, Raquel O. Prates, Fabrício Benevenuto, and Virgílio Almeida</i>	

Sound and Smell

The Role of Music in the Design Process with Children	288
<i>Ruut Tikkanen and Netta Iivari</i>	

ToCoPlay: Graphical Multi-touch Interaction for Composing and Playing Music	306
<i>Sean Lynch, Miguel A. Nacenta, and Sheelagh Carpendale</i>	

Presentation Technique of Scents Using Mobile Olfactory Display for Digital Signage	323
<i>Sayumi Sugimoto, Ryo Segawa, Daisuke Noguchi, Yuichi Bannai, and Kenichi Okada</i>	

Touch Interfaces

“Oh Snap” – Helping Users Align Digital Objects on Touch Interfaces	338
<i>Jennifer Fernquist, Garth Shoemaker, and Kellogg S. Booth</i>	

The Link-Offset-Scale Mechanism for Improving the Usability of Touch Screen Displays on the Web	356
<i>William Massami Watanabe, Renata Pontin de Mattos Fortes, and Maria da Graça Campos Pimentel</i>	

The Effects of Personal Displays and Transfer Techniques on Collaboration Strategies in Multi-touch Based Multi-Display Environments	373
<i>Stefan Bachl, Martin Tomitsch, Karin Kappel, and Thomas Grechenig</i>	

Tabletops I

Evaluating Physical/Virtual Occlusion Management Techniques for Horizontal Displays	391
<i>Waqas Javed, KyungTae Kim, Sohaib Ghani, and Niklas Elmqvist</i>	

Usage and Recognition of Finger Orientation for Multi-Touch Tabletop Interaction 409
Chi Tai Dang and Elisabeth André

Tangoscope: A Tangible Audio Device for Tabletop Interaction 427
Jörg Edelmann, Yvonne Kammerer, Birgit Imhof, Peter Gerjets, and Wolfgang Straßer

Supporting Social Protocols in Tabletop Interaction through Visual Cues 435
Mirko Fetter, Tom Gross, and Maxi Hucke

Tabletops II

Effects of a Tabletop Interface on the Co-construction of Concept Maps 443
Stefan Oppl and Chris Stary

The Continuous Interaction Space: Interaction Techniques Unifying Touch and Gesture on and above a Digital Surface 461
Nicolai Marquardt, Ricardo Jota, Saul Greenberg, and Joaquim A. Jorge

AffinityTable - A Hybrid Surface for Supporting Affinity Diagramming 477
Florian Geyer, Ulrike Pfeil, Jochen Budzinski, Anita Höchtl, and Harald Reiterer

Ubiquitous and Context-Aware Computing

Design as Intercultural Dialogue: Coupling Human-Centered Design with Requirement Engineering Methods 485
Chiara Leonardi, Luca Sabatucci, Angelo Susi, and Massimo Zancanaro

Predicting Selective Availability for Instant Messaging 503
Mirko Fetter, Julian Seifert, and Tom Gross

Testing the Usability of a Platform for Rapid Development of Mobile Context-Aware Applications 521
Valentim Realinho, A. Eduardo Dias, and Teresa Romão

UI Modeling I

Hammering Models: Designing Usable Modeling Tools 537
Ko-Hsun Huang, Nuno Jardim Nunes, Leonel Nobrega, Larry Constantine, and Monchu Chen

Task Descriptions Using Academic Oriented Modelling Languages: A Survey of Actual Practices across the SIGCHI Community	555
<i>Stanislas Cowix and Jean-Marie Burkhardt</i>	
Selective Modeling to Support Task Migratability of Interactive Artifacts	571
<i>Anke Dittmar and Peter Forbrig</i>	
UI Modelling II	
Structuring and Composition Mechanisms to Address Scalability Issues in Task Models	589
<i>Célia Martinie, Philippe Palanque, and Marco Winckler</i>	
User Driven Evolution of User Interface Models – The FLEPR Approach	610
<i>Stefan Hennig, Jan Van den Bergh, Kris Luyten, and Annerose Braune</i>	
Adapting Desktop Web Pages for Vocal Browsing	628
<i>Fabio Paternò and Christian Sisti</i>	
Using the Journalistic Metaphor to Design User Interfaces That Explain Sensor Data	636
<i>Martin Molina, Enrique Parodi, and Amanda Stent</i>	
Usability	
Domain Experts Tailoring Interaction to Users – An Evaluation Study	644
<i>Helena Lindgren, Patrik J. Winnberg, and Peter Winnberg</i>	
Identifying Relationships between Physiological Measures and Evaluation Metrics for 3D Interaction Techniques	662
<i>Rafael Rieder, Christian Haag Kristensen, and Márcio Sarroglia Pinho</i>	
Comparing User Experience and Performance in SecondLife and Blackboard	680
<i>Alistair G. Sutcliffe and Amal Alrayes</i>	
Author Index	697

Table of Contents – Part IV

Long and Short Papers

Usable Privacy and Security

A Field Study of User Behavior and Perceptions in Smartcard Authentication	1
<i>Celeste Lyn Paul, Emile Morse, Aiping Zhang, Yee-Yin Choong, and Mary Theofanos</i>	
Improving Computer Security Dialogs	18
<i>Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, Saranga Komanduri, and Manya Sleeper</i>	
Usable Privacy and Security in Personal Health Records	36
<i>Inma Carrión, Jose L. Fernández-Alemán, and Ambrosio Toval</i>	
Shining Chrome: Using Web Browser Personas to Enhance SSL Certificate Visualization	44
<i>Max-Emanuel Maurer, Alexander De Luca, and Tobias Stockinger</i>	

User Experience I

Ambient Displays: Influencing Movement Patterns	52
<i>Tasos Varoudis</i>	
Three User-Driven Innovation Methods for Co-creating Cloud Services	66
<i>Ting-Ray Chang and Eija Kaasinen</i>	
Designing for the Secondary User Experience	84
<i>Ole Andreas Alsos and Dag Svanæs</i>	
Engaging Visitors in Museums with Technology: Scales for the Measurement of Visitor and Multimedia Guide Experience	92
<i>Mohd Kamal Othman, Helen Petrie, and Christopher Power</i>	

User Experience II

An Image of Electricity: Towards an Understanding of How People Perceive Electricity	100
<i>Yoram Chisik</i>	

Structuring the Collaboration of Multiple Novice Design Ethnographers:
Towards a New User Research Approach 118
*Paul Gault, Catriona Macaulay, Graham Johnson, and
Judith Masthoff*

Customer Experience Modeling: Designing Interactions for Service
Systems 136
Jorge Teixeira, Lia Patrício, Nuno J. Nunes, and Leonel Nóbrega

User Experience Research in the Semiconductor Factory:
A Contradiction? 144
*Marianna Obrist, Wolfgang Reitberger, Daniela Wurhofer,
Florian Förster, and Manfred Tscheligi*

User Experience III

Client’s Temporal Trajectory in Child Protection: Piecing Information
Together in a Client Information System 152
Saila Huuskonen and Pertti Vakkari

Unsupervised Parameter Selection for Gesture Recognition with Vector
Quantization and Hidden Markov Models 170
*Przemysław Głomb, Michał Romaszewski, Arkadiusz Sochan, and
Sebastian Opozda*

Number Entry Interfaces and Their Effects on Error Detection 178
Patrick Oladimeji, Harold Thimbleby, and Anna Cox

An Exploration of the Utilization of Electroencephalography and
Neural Nets to Control Robots 186
Dan Szafir and Robert Signorile

Social Translucence as a Theoretical Framework for Sustainable HCI ... 195
Mary Barreto, Evangelos Karapanos, and Nuno Nunes

User Modelling

A Revised Mobile KLM for Interaction with Multiple NFC-Tags 204
Paul Holleis, Maximilian Scherr, and Gregor Broll

The Entropy of a Rapid Aimed Movement: Fitts’ Index of Difficulty
versus Shannon’s Entropy 222
R. William Soukoreff, Jian Zhao, and Xiangshi Ren

The Difference Matters: Benchmarking Visual Performance of a
Cognitive Pilot Model 240
Florian Frische and Andreas Lüdtke

Visual Search in Radial Menus	248
<i>Krystian Samp and Stefan Decker</i>	

Visualization

Analytic Trails: Supporting Provenance, Collaboration, and Reuse for Visual Data Analysis by Business Users	256
<i>Jie Lu, Zhen Wen, Shimei Pan, and Jennifer Lai</i>	

Exploration Views: Understanding Dashboard Creation and Customization for Visualization Novices	274
<i>Micheline Elias and Anastasia Bezerianos</i>	

Patient Development at a Glance: An Evaluation of a Medical Data Visualization	292
<i>Margit Pohl, Sylvia Wiltner, Alexander Rind, Wolfgang Aigner, Silvia Miksch, Thomas Turic, and Felix Drexler</i>	

Evaluation of HaloDot: Visualization of Relevance of Off-Screen Objects with over Cluttering Prevention on Mobile Devices	300
<i>Tiago Gonçalves, Ana Paula Afonso, Maria Beatriz Carmo, and Paulo Pombinho</i>	

Web Interaction

Using Card Sorts for Understanding Website Information Architectures: Technological, Methodological and Cultural Issues	309
<i>Helen Petrie, Christopher Power, Paul Cairns, and Cagla Seneler</i>	

The Treatment of Temporal Data in Web-Based Reservation Systems: An Inspection-Based Evaluation	323
<i>Gerhard F. Knolmayer, Lukas E. Helfenstein, and Viola Sini</i>	

A Tool Support for Web Applications Adaptation Using Navigation History	340
<i>Sergio Firmenich, Marco Winckler, and Gustavo Rossi</i>	

Web Usability Probe: A Tool for Supporting Remote Usability Evaluation of Web Sites	349
<i>Tonio Carta, Fabio Paternò, and Vagner Figuerêdo de Santana</i>	

Demos

A Demo of a Dynamic Facial UI for Digital Artists	358
<i>Pedro Bastos, Xenxo Alvarez, and Veronica Orvalho</i>	

A Set of Customizable Games Supporting Therapy of Children with Cerebral Palsy	360
<i>Beant Dhillon, Areti Goulati, Ioannis Politis, Agata Raczevska, and Panos Markopoulos</i>	
Mobile Total Conversation – Communication for All, Everywhere	362
<i>Erik Zetterström</i>	
Storytelling Meets the Social Web: An HTML5 Cross-Platform Application for Older Adults	364
<i>Tiago Boldt Sousa, Pedro Tenreiro, Paula Alexandra Silva, and Eduarda Mendes Rodrigues</i>	
Tablexcel: A Multi-user, Multi-touch Interactive Tabletop Interface for Microsoft Excel Spreadsheets	366
<i>Guillaume Besacier</i>	
Doctoral Consortium	
Assessing Short-Term Human-Robot Interaction in Public Space	370
<i>Jakub Zlotowski</i>	
Barefooted Usability Evaluation: Addressing the Mindset, Resources and Competences	374
<i>Anders Bruun</i>	
Collaborative Human-Machine Communication: User-Centered Design of In-Vehicle Speech Dialog Systems	378
<i>Linn Hackenberg</i>	
Development of a Methodology for Evaluating the Quality in Use of Web 2.0 Applications	382
<i>Tihomir Orehovački</i>	
Distortion Techniques for Sketching Interaction	386
<i>Paul Schmieder</i>	
Evaluation of Information Classification on Websites and Impact of Culture: A Cross Country Comparison of Information Classification	390
<i>Ather Nawaz</i>	
Exploring New Ways of Utilizing Automated Clustering and Machine Learning Techniques in Information Visualization	394
<i>Johann Schrammel</i>	
Exploring Serendipity’s Precipitating Conditions	398
<i>Lori McCay-Peet</i>	

Human-Computer Interaction and Human Mental Workload: Assessing Cognitive Engagement in the World Wide Web	402
<i>Luca Longo</i>	
Human-Computer Interaction for Security Research: The Case of EU E-Banking Systems	406
<i>Caroline Moeckel</i>	
Information Architecture Automatization for the Semantic Web	410
<i>Josep Maria Brunetti and Roberto García</i>	
Microinteractions to Augment Manual Tasks	414
<i>Katrin Wolf</i>	
OPEN-HEREDEUX: OPEN HEuristic REsource for Designing and Evaluating User eXperience	418
<i>Llúcia Masip, Marta Oliva, and Toni Granollers</i>	
Sketching Language: User-Centered Design of a Wizard of Oz Prototyping Framework	422
<i>Stephan Schlögl</i>	
Time Affordances and Physical Mobility in the Context of Ubiquitous Technologies	426
<i>Larissa Pschetz</i>	
Usability Evaluation in Software Development Practice	430
<i>Marta Kristin Larusdottir</i>	
Website Customization: Exploring a Tag-Based Approach in the Australian Banking Context	434
<i>Rajinesh Ravendran</i>	

Industrial Papers

Acceptance and Speed of Animations in Business Software	438
<i>Lisa Mattes, Martin Schrepp, Theo Held, and Patrick Fischer</i>	
Developing Mobile Remote Collaboration Systems for Industrial Use: Some Design Challenges	442
<i>Leila Alem and Weidong Huang</i>	
Experiences of Online Co-creation with End Users of Cloud Services	446
<i>Kaarina Karppinen, Kaisa Koskela, Camilla Magnusson, and Ville Nore</i>	
Interactive Installations: Tales from the Trenches	450
<i>Pedro Campos, Miguel Campos, and Joaquim A. Jorge</i>	

Interactive Posters

A Conceptual Framework for Modeling Awareness Mechanisms in Collaborative Systems	454
<i>Fernando Gallego, Ana Isabel Molina, Jesús Gallardo, and Crescencio Bravo</i>	
A Longitudinal Pilot Study to Evaluate Non-visual Icons in a Mobile Exertion Application	458
<i>Huimin Qian, Ravi Kuber, and Andrew Sears</i>	
A Remote Multi-touch Experience to Support Collaboration between Remote Museum Visitors	462
<i>Ernesto Arroyo, Valeria Righi, Roger Tarrago, and Josep Blat</i>	
A Systematic Evaluation of Mobile Applications for Diabetes Management	466
<i>C. Martin, D. Flood, D. Sutton, A. Aldea, R. Harrison, and M. Waite</i>	
An Integrated Approach to Develop Interactive Software	470
<i>Begoña Losada, Maite Urretavizcaya, and Isabel Fernández de Castro</i>	
Analyzing the Level of Presence While Navigating in a Virtual Environment during an fMRI Scan	475
<i>Miriam Clemente, Alejandro Rodríguez, Beatriz Rey, Aina Rodríguez, Rosa M. Baños, Cristina Botella, Mariano Alcañiz, and César Ávila</i>	
Applying the Affinto Ontology to Develop a Text-Based Emotional Conversation System	479
<i>Idoia Cearreta and Nestor Garay</i>	
Augmented Mirror: Interactive Augmented Reality System Based on Kinect	483
<i>Lucía Vera, Jesús Gimeno, Inmaculada Coma, and Marcos Fernández</i>	
Calls for Interaction: The More the Better? User Experience of 3D Carousel and Additional Interaction Techniques	487
<i>S. Shyam Sundar, Saraswathi Bellur, Jeeyun Oh, and Haiyan Jia</i>	
Can Persona Facilitate Ideation? A Comparative Study on Effects of Personas in Brainstorming	491
<i>Xiantao Chen, Ying Liu, Ning Liu, and Xiaojie Wang</i>	
Children with Special Needs: Comparing Tactile and Tangible Interaction	495
<i>César Ortea Suárez, Javier Marco, Sandra Baldassarri, and Eva Cerezo</i>	

Coupling Interaction and Physiological Metrics for Interaction Adaptation	499
<i>Luís Duarte and Luís Carriço</i>	
Dual Flow Interaction: Scene Flow and Data Flow, Dual Interaction in Art Installations	503
<i>José M^a Alonso-Calero, Arcadio Reyes-Lecuona, Jesus Marín-Clavijo, and Josefa Cano-García</i>	
Effects of Touch Screen Response Time on Psychological State and Task Performance	507
<i>Nozomi Sato and Kentaro Nakajima</i>	
Elaborating Analysis Models with Tool Support	511
<i>Gregor Buchholz and Peter Forbrig</i>	
End-User Support for Information Architecture Analysis in Interactive Web Applications	515
<i>Luis A. Rojas and José A. Macías</i>	
Enriching Evaluation in Video Games	519
<i>José L. González Sánchez, Rosa M. Gil Iranzo, and Francisco L. Gutiérrez Vela</i>	
Evaluation of 3D Object Manipulation on Multi-touch Surfaces Using Unconstrained Viewing Angles	523
<i>Daniel Mendes and Alfredo Ferreira</i>	
Evaluation of an Accessible Home Control and Telecare System	527
<i>Fausto Sainz, Judit Casacuberta, Marta Díaz, and Jaisiel Madrid</i>	
Experimenting and Improving Perception of 3D Rotation-Based Transitions between 2D Visualizations	531
<i>Maxime Cordeil, Christophe Hurter, and Stéphane Conversy</i>	
HTML 5 Support for an Accessible User-Video-Interaction on the Web	535
<i>Lourdes Moreno, Paloma Martínez, Ana Iglesias, and María Gonzalez</i>	
Improving the Flexibility of Model Transformations in the Model-Based Development of Interactive Systems	540
<i>Christian Wiehr, Nathalie Aquino, Kai Breiner, Marc Seissler, and Gerrit Meixner</i>	
In Two Minds about Usability? Rationality and Intuition in Usability Evaluations	544
<i>Volker Thoma and Elliott P. White</i>	

Influence of Web Content Management Systems in Web Content Accessibility	548
<i>Juan Miguel López, Afra Pascual, Lucía Masip, Toni Granollers, and Xavier Cardet</i>	
Instructional Animations: More Complex to Learn from Than at First Sight?	552
<i>Anna Wong, Nadine Marcus, and John Sweller</i>	
Integrating Feedback into Wearable Controls	556
<i>Cátia Sousa and Ian Oakley</i>	
Intelligent Playgrounds: Measuring and Affecting Social Inclusion in Schools	560
<i>Olga Lyra, Evangelos Karapanos, and Vassilis Kostakos</i>	
It Does Not Fitts My Data! Analysing Large Amounts of Mobile Touch Data	564
<i>Niels Henze and Susanne Boll</i>	
Measuring Cognitive Workload with Low-Cost Electroencephalograph	568
<i>Avi Knoll, Yang Wang, Fang Chen, Jie Xu, Natalie Ruiz, Julien Epps, and Pega Zarjam</i>	
Model-Based Accessible User Interface Generation in Ubiquitous Environments	572
<i>Raúl Miñón, Julio Abascal, Amaia Aizpurua, Idoia Cearreta, Borja Gamecho, and Nestor Garay</i>	
Multiuser Augmented Reality System for Indoor Exhibitions	576
<i>Jesus Gimeno, Ricardo Olanda, Bibiana Martinez, and Fernando M. Sanchez</i>	
Natural Interaction without Marks	580
<i>Carina S. González-González, David Cabrera-Primo, Melvin Gutierrez, and Jose Sigut-Saavedra</i>	
NAVI – A Proof-of-Concept of a Mobile Navigational Aid for Visually Impaired Based on the Microsoft Kinect	584
<i>Michael Zöllner, Stephan Huber, Hans-Christian Jetter, and Harald Reiterer</i>	
OntoCompo: A Tool to Enhance Application Composition	588
<i>Christian Brel, Anne-Marie Dery-Pinna, Philippe Renevier-Gonin, and Michel Riveill</i>	

Personal Chart: Health Data Logging Made Easy with a Mobile Device	592
<i>Mikko Paldanius, Anu Lehtiö, Minna Karukka, and Pertti Huuskonen</i>	
Psychosocial Indicators via Hand Tremor	596
<i>Ted Selker, Patricia Collins, and Will Dayton</i>	
Recognizing Emotions from Video in a Continuous 2D Space	600
<i>Sergio Ballano, Isabelle Hupont, Eva Cerezo, and Sandra Baldassarri</i>	
Supporting Moodle-Based Lesson through Visual Analysis	604
<i>Diego Gomez-Aguilar, Miguel Conde-Gonzalez, Roberto Theron, and Francisco Garcia-Peñalvo</i>	
Supporting Transformations across User Interface Descriptions at Various Abstraction Levels	608
<i>Mauro Lisai, Fabio Paternò, Carmen Santoro, and Lucio Davide Spano</i>	
Texture Recognition: Evaluating Force, Vibrotactile and Real Feedback	612
<i>Jonatan Martínez, Arturo S. García, Diego Martínez, José P. Molina, and Pascual González</i>	
The Application of Preference Mapping in Aesthetic Website Evaluation	616
<i>Eleftherios Papachristos and Nikolaos Avouris</i>	
The Effect of Religious Identity on User Judgment of Website Quality	620
<i>Ons Al-shamaileh, Alistair Sutcliffe, and Antonella De Angeli</i>	
Toward a Better Guidance in Wearable Electronic Orientation Aids	624
<i>Slim Kammoun, Marc J.-M. Macé, Bernard Oriola, and Christophe Jouffrais</i>	
Towards a Context Oriented Approach to Ethical Evaluation of Interactive Technologies	628
<i>Sandra Burri Gram-Hansen, Henrik Schärfe, and Jens Vilhelm Dinesen</i>	
Towards a Framework of Co-Design Sessions with Children	632
<i>Emanuela Mazzone, Janet C. Read, and Russell Beale</i>	
Towards a Semantic Modelling Framework in Support of Multimodal User Interface Design	636
<i>Elena Tsiorkova, Tom Tourwé, and Nicolás González-Deleito</i>	

Towards an Experimental Framework for Measuring Usability of Model-Driven Tools	640
<i>Jose Ignacio Panach, Nelly Condori-Fernández, Arthur Baars, Tanja Vos, Ignacio Romeu, and Óscar Pastor</i>	
TROCAS: Communication Skills Development in Children with Autism Spectrum Disorders via ICT	644
<i>Margarida Lucas da Silva, Carla Simões, Daniel Gonçalves, Tiago Guerreiro, Hugo Silva, and Fernanda Botelho</i>	
Usability Assessment of a Multimodal Visual-Haptic Framework for Chemistry Education	648
<i>Sara Comai and Davide Mazza</i>	
Usability Planner: A Tool to Support the Process of Selecting Usability Methods	652
<i>Xavier Ferre and Nigel Bevan</i>	
User Experience Specification through Quality Attributes	656
<i>Llúcia Masip, Marta Oliva, and Toni Granollers</i>	
Using Availability Heuristics in Game Design to Introduce Children to Energy Sufficient Behaviours at Home	661
<i>Nsemeke Ukpog, Privender Saini, and Abdullah Al Mahmud</i>	
UsiXML Extension for Awareness Support	665
<i>Jose Figueroa-Martínez, Francisco L. Gutiérrez Vela, Víctor López-Jaquero, and Pascual González</i>	
Web Accessibility Requirements for Media Players	669
<i>María González, Lourdes Moreno, Paloma Martínez, and Ana Iglesias</i>	
Organization Overviews	
Christian Doppler Laboratory: Contextual Interfaces	675
<i>David Wilfinger, Alexander Meschtscherjakov, Astrid Weiss, and Manfred Tscheligi</i>	
Interaction Modeling at PROS Research Center	677
<i>José Ignacio Panach, Nathalie Aquino, and Oscar Pastor</i>	
Overview of the Brazilian Computer Society's Council for Human-Computer Interaction (CEIHC)	679
<i>Cristiano Maciel, Elizabeth Furtado, Marco Winckler, Milene Silveira, and Raquel Prates</i>	

Supporting a Multidisciplinary Digital Media Research Community with GRAND Aspirations	681
<i>Kellogg S. Booth and Eleni Stroulia</i>	

The Centre for Internationalization and Usability: Enabling Culture-Centred Design for All	683
<i>José Abdelnour-Nocera, Andy Smith, John Moore, Cecilia Oyugi, Souleymane Camara, Malte Ressin, Sujan Shresta, and Alison Wiles</i>	

Panels

Critical Design :: Is It Just Designers Doing Ethnography or Does It Offer Something More for Interaction Design?	685
<i>Michael Smyth, Chris Speed, and Martin Brynskov</i>	

Everyone is a Designer, Even Executives!	687
<i>Jannie Lai and Iram Mirza</i>	

Special Interest Groups (SIGs)

HCI for Peace: Promoting Peace and Preventing War through Computing Technology	689
<i>Juan Pablo Hourcade, Natasha E. Bullock-Rest, Janet C. Read, and Yoram Chisik</i>	

Interaction and Music Technology	691
<i>Sidney Fels and Michael Lyons</i>	

User Interface eXtensible Markup Language SIG	693
<i>Gaëlle Calvary, Olivier de Wasseige, David Faure, and Jean Vanderdonckt</i>	

Tutorials

Activity-Centered Interaction Design: A Model-Driven Approach	696
<i>Larry Constantine</i>	

Analysis, Redesign and Evaluation with Teasing Apart, Piecing Together	698
<i>Clare J. Hooper</i>	

Context-Aware Adaptation of User Interfaces	700
<i>Vivian Genaro Motti and Jean Vanderdonckt</i>	

Designing the Search Experience	702
<i>Tony Russell-Rose</i>	

Improving the Content of User Requirements	704
<i>Nigel Bevan</i>	
Model-Driven Inquiry: Beyond Ethnography and Contextual Inquiry . . .	706
<i>Larry Constantine</i>	
Scenario-Based Requirements Engineering Facilitating Interaction Design	708
<i>Hermann Kaindl</i>	
Sketching Interactive Systems with Sketchify	710
<i>Željko Obrenović</i>	
UIs Automatically Optimized for Your Smartphone	712
<i>Hermann Kaindl</i>	
User Experience Evaluation – Which Method to Choose?	714
<i>Virpi Roto, Arnold Vermeeren, Kaisa Väänänen-Vainio-Mattila, and Effie Law</i>	
User Experience Evaluation in Entertainment and Games	716
<i>Regina Bernhaupt</i>	

Workshops

5 th Workshop on Software and Usability Engineering Cross-Pollination: Patterns, Usability and User Experience	718
<i>Peter Forbrig, Regina Bernhaupt, Marco Winckler, and Janet Wesson</i>	
Accessible Design in the Digital World	720
<i>Gerhard Weber, Helen Petrie, and Jenny Darzentas</i>	
Building Bridges – HCI and Visualization	722
<i>Achim Ebert, Gitta Domik, Nahum Gershon, and Gerrit van der Veer</i>	
Combining Design and Engineering of Interactive Systems through Models and Tools (ComDeisMoto)	724
<i>Stefan Sauer, Kai Breiner, Heinrich Hussmann, Gerrit Meixner, Andreas Pleuss, and Jan Van den Bergh</i>	
Data-Centric Interactions on the Web	726
<i>Paloma Díaz, Tim Hussein, Steffen Lohmann, and Jürgen Ziegler</i>	
Encouraging Serendipity in Interactive Systems	728
<i>Stephann Makri, Elaine G. Toms, Lori McCay-Peet, and Ann Blandford</i>	

Human Work Interaction Design for e-Government and Public Information Systems	730
<i>Dinesh Katre, Pedro Campos, Torkil Clemmensen, Rikke Orngreen, and Annelise Mark Pejtersen</i>	
Improving the Content of User Requirements	732
<i>Nigel Bevan</i>	
Mobile Accessibility Workshop	734
<i>Daniel Gonçalves, Luis Carriço, and Markel Vigo</i>	
Promoting and Supporting Healthy Living by Design	736
<i>Gordon Baxter, Lisa Dow, Stephen Kimani, and Nilufar Baghaei</i>	
Re-framing HCI through Local and Indigenous Perspectives	738
<i>Jose Abdelnour-Nocera, Masaaki Kurosu, Torkil Clemmensen, Nic Bidwell, Ravi Vatrapu, Heike Winschiers-Theophilus, Vanessa Evers, Rüdiger Heimgärtner, and Alvin Yeo</i>	
Software Support for User Interface Description Language	740
<i>Adrien Coyette, David Faure, Juan González-Calleros, and Jean Vanderdonckt</i>	
User Experience in Cars	742
<i>Manfred Tscheligi, Albrecht Schmidt, David Wilfinger, Alexander Meschtscherjakov, and Andrew L. Kun</i>	
User Interaction Techniques for Future Lighting Systems	744
<i>Dzmitry Aliakseyeu, Jon Mason, Bernt Meerbeek, Harm van Essen, Serge Offermans, and Andrés Lucero</i>	
Values in Design - Building Bridges between RE, HCI and Ethics	746
<i>Christian Detweiler, Alina Pommeranz, Jeroen v.d. Hoven, and Helen Nissenbaum</i>	
Author Index	749

Natural User Interfaces

António Câmara

YDreams - Informática S.A.

antonio.camara@ydreams.com

<http://www.ydreams.com/>

Abstract. Recent developments in user-input technologies are changing the way we interact with digital screens. The mouse and the keyboard are being replaced by touch and motion based interfaces, increasingly known as Natural User Interfaces (NUI). YDreams has developed Yvision, a platform that enables the development of natural user interfaces. YVision has a modular architecture matching YDreams technologies with the best of open source third party libraries. Our technologies emphasize the creation of smart interfaces using autonomous agents that go beyond the traditional reactive systems. Yvision also includes computer vision algorithms for motion detection and the application of 3D depth sensing in rendering engines. NUI applications involve: data acquisition using various sensors that detect the user's motion and gestures; interpretation of sensor data; and presentation, the end visualization layer. YVision includes augmented reality capabilities as a visualization component, where images are captured from the real world and enhanced in real-time with contextual information. Natural user interface applications, developed for both 2D and 3D depth sensing, will be presented for illustrative purposes. Applications include projects developed for clients such as Orange, Coca-Cola, Santander and Nike. Ongoing research projects focusing on digital signage and serious games will be also discussed.

Short Biography

António Câmara is Chief Executive Officer of YDreams and Professor at Universidade Nova de Lisboa. He got a BSc in Civil Engineering at IST (1977) and MSc (1979) and PhD (1982) in Environmental Systems Engineering at Virginia Tech. António Câmara was a Post-Doctoral Associate at Massachusetts Institute of Technology (MIT) and Visiting Professor at Cornell University (1988-89) and MIT (1998-99). António Câmara has been a pioneer on geographical information systems research. He published over 150 refereed papers and the "Spatial Multimedia and Virtual Reality" published by Taylor & Francis (1999) and "Environmental Systems" published by Oxford University Press (2002). He is a founder of YDreams, a international leader in interactivity. YDreams has developed more than 600 projects in 25 countries for companies such as Nike, Adidas, Santander, Coca-Cola, NOKIA and Vodafone. The company has received over twenty awards including the Industrial Design Society of America Gold Award for Interactive Environments in 2004, and the Auggies, Augmented Reality's Oscar, in 2010. YDreams projects and products have been profiled in the New York Times, Guardian, Liberation, El Pais, Business Week, Economist, Wired, Engadget, Gizmodo, CNN and CNBC. António Câmara has received several national and international awards, namely Premio Pessoa in 2006.

The Future of Distributed Groups and Their Use of Social Media

Mary Czerwinski

Microsoft Research
Visualization and Interaction (VIBE) Research Group
marycz@microsoft.com

Abstract. Distributed team field research has shown that shared group awareness, coordination and informal communication are the most common ways for teams to inform each other of progress. In addition, we have observed that poorly documented, informal communication causes a fragmented workday due to frequent interruptions and knowledge loss due to the passage of time and team attrition. Because informal communication has both advantages and disadvantages for information sharing, it merits deeper study to allow any proposed solution to preserve the good while reducing the bad. Over the past several years, we have conducted a series of studies at Microsoft Corporation and beyond to document the nature of group conversations and communications. Based on surveys, lab studies, field studies and interviews, we have begun to develop a suite of tools that allow groups, both co-located and distributed, to stay more aware of their colleagues' actions, get on board to a new team more efficiently, and engage with each other at the most optimal times. Examples of many of these tools will be discussed, as will our progress in transitioning these ideas into real products.

Short Biography

Bio: Mary Czerwinski is a Research Area Manager at Microsoft Research, where she manages many diverse areas of human-computer interaction, including social computing, information visualization, CSCW, sensor-based interaction and healthcare. Mary has been an avid participant in the ACM SIGCHI community, sitting on the SIGCHI Executive Committee for the last 10 years, chairing CHI 2008, UIST 2005, Papers Chair for CHI 2000 and UIST 2010, in addition to many other conference volunteer roles. Mary was recently awarded the ACM SIGCHI Lifetime Service award and was also inducted into the ACM CHI Academy. Mary has ~100 publications in HCI and psychology, and holds a PhD in Cognitive Psychology. Mary is very involved in supporting academia as well, sitting on multiple university advisory boards and PhD student dissertation committees.

Opportunities for Proxemic Interactions in Ubicomp (Keynote)

Saul Greenberg

Department of Computer Science, University of Calgary, Calgary, AB, Canada T2N 1N4
saul.greenberg@ucalgary.ca

Abstract. In this keynote presentation, I describe and illustrate *proxemic interactions* as realized in several projects in my laboratory. My goal is to advocate proxemics as a more natural way of mediating inter-entity interactions in ubiquitous computing environments, while still cautioning about the many pitfalls around its use.

Keywords: Proxemic interactions, ubiquitous computing, interaction techniques, ubicomp ecologies.

1 Introduction

In the everyday world, much of what we do as social beings is dictated by how we interpret spatial relationships. This is called *proxemics*, and is part of the glue that comprises our natural social fabric. What is surprising is how little people's expectations of spatial relationships are used in interaction design within ubiquitous computing – ubicomp [1] – ecologies, i.e., in terms of mediating people's interactions with surrounding digital devices such as digital surfaces, mobile phones, tablets, computers, and other information appliances. *Proxemic interaction* imagines a world of devices that have fine-grained knowledge of nearby people and other devices – how they move into range, their precise distance, their identity, their orientation and their context – and how such knowledge can be exploited to design interaction techniques. Just as people expect increasing engagement and intimacy as they approach others, so should they naturally expect increasing connectivity and interaction possibilities as they bring themselves and their devices in close proximity to one another and to other things in their everyday ecology.

This keynote describes and illustrates proxemic interactions¹. My goal is to advocate proxemics as a more natural way of mediating inter-entity interactions in ubicomp environments, while still cautioning about the many pitfalls around its use.

2 What Is Proxemics?

Anthropologist Edward Hall [7] introduced *proxemics* as an area of study that identifies ways that people use inter-personal distance to understand and mediate their

¹ The keynote address and this article is largely based upon [2], with many inspirations drawn from [3-6] as well as the myriad of wonderful works done by other researchers.

interactions with other people. He was primarily concerned with culturally-dependent attributes of proxemics, where (for example) people from a Western society would consider inter-personal distance somewhat differently from people in an Arabic Society. Along the way, he developed a theory of proxemics that articulates how different things affect people's perception of inter-personal distance.

First are his definitions of four *proxemic 'zones'*, which correlate physical to social distance. In essence, the closer the distance the more it leads to increasing expectations (and perhaps violations if the social conditions don't warrant it) of inter-personal engagement and intimacy. A second factor that influences people's proxemics interactions are the location of *fixed-* and *semi-fixed features* within the space. Fixed features include those that mark boundaries (e.g., entrances to a particular type of room), where people tend to organize certain kinds of social activities within these boundaries. Semi-fixed features are entities whose position can affect whether the space tends to bring people together, or move them apart e.g., the placement of furniture. Of course, many other factors influence proxemics. These include the social relations of the people involved, gaze and eye contact, voice volume, posture, body language, cultural taboos about touch, contextual factors such as crowding, and so on.

3 How Can We Adapt Proxemics to Ubicomp?

Now comes the speculation: Can we apply the theory of proxemics to ubicomp? I restate this as a working hypothesis:

Just as people expect increasing engagement and intimacy as they approach others, so should they naturally expect increasing connectivity and interaction possibilities as they bring themselves and their devices in close proximity to each other and to other things in the ecology.

The caveat is that proxemics, as envisioned by Hall and others that followed, never included ubicomp as part of its theory. Proxemics was about people's relationship to people, not to devices. Even so, I believe that proxemics can be applied as a first-order approximation that characterizes not only people to people, but people to device and device to device inter-entity relationships and expectations. This is not as far-fetched as it seems, especially because sociologists have found that people often react to computational devices as social entities, e.g., [8].

Our first challenge is to operationalize the concept of proximity in ubicomp. For devices to react, devices must be able to capture and use information from the environment as variables. While there are likely many variables we could use, we previously identified 5 variables – which we call proxemic dimensions – as a starting point [2].

- **Distance** between entities is fundamental. It can be captured as a continuous measure (e.g., a value between 0 - 6 feet), as a discrete measure of what zone an entity is in with respect to another entity (e.g., [9, 10]), or even as just a binary measure, e.g., one entity is or is not in the same room as another entity.
- **Orientation** captures the way one entity is facing another (e.g., [11]). It too can be precise and continuous (e.g., the exact pitch/roll/yaw angle of one entity relative to another), or discrete (e.g., facing towards, somewhat towards, or away

from the other object). Of course, orientation only makes sense if an entity has a ‘front face’ to it.

- **Identity** uniquely describes the entity. This can include exact identity and attributes, to a less detailed measure such as an entity’s type, to a minimal measure that simply discriminates one entity from another.
- **Movement** captures the distance and orientation of an entity over time, where different actions can be taken depending on (for example) the speed of motion, and/or whether one entity is moving and turning towards vs. away from another entity.
- **Location** describes the physical context that the entities reside in. It can capture contextual aspects, such as when an entity crosses a threshold (a fixed feature) marking its presence or departure from a room. Crucially, the meaning applied to the four other inter-entity measures may depend on the contextual location, for example, what ‘rules’ get fired in different locations.

Each measure has appeared before in other ubicomp systems, and various researchers have applied one or more of these to proximity (e.g., [9-19]). Yet very few make use of all of them, let alone consider them as characterizing the interplay between entities in an ubicomp ecology. One of the hurdles is that most developers of these kinds of systems have to construct all the low-level sensing from scratch, which can significantly hinder prototype development and iteration [20]. To mitigate this problem, we developed the Proximity Toolkit [21, 6], a toolkit that tracks the above variables and presents them to the programmer both visually and through an API that characterizes the relationship between entities.

4 Vignettes of What We Can Do with Proxemics

Our efforts in prototyping software in our laboratory, as well as efforts by other researchers who have worked in the area, have produced many different examples of how knowledge of proxemics can be applied to interaction design. I characterize some of the things that have been done to illustrate the many opportunities available, via a scenario of vignettes based on the ubicomp ecology – a room – illustrated in Fig. 1. Each vignette is preceded by a heading that describes how proxemics is used to advantage. I don’t go into any interface specifics here, but the references included provide pointers to particular systems that exploit and implement these strategies.

Power management and sustainability. John enters the room. Both the large display on the wall and the picture frame on the table (the information appliance) recognize that a person has crossed the threshold (the fixed feature) into the room. Each powers itself up from sleep mode so as to be readily available. When John leaves, they immediately go back to sleep. [20]

Initial state / Ambient display. Each display shows information appropriate for a just-entering person who may not be directly attending the display. In this case, the wall display shows the media player John had on when he was last in the room. It highlights – using large images and minimal text appropriate for a quick glance at a distance – a few videos that he was considering watching [4].

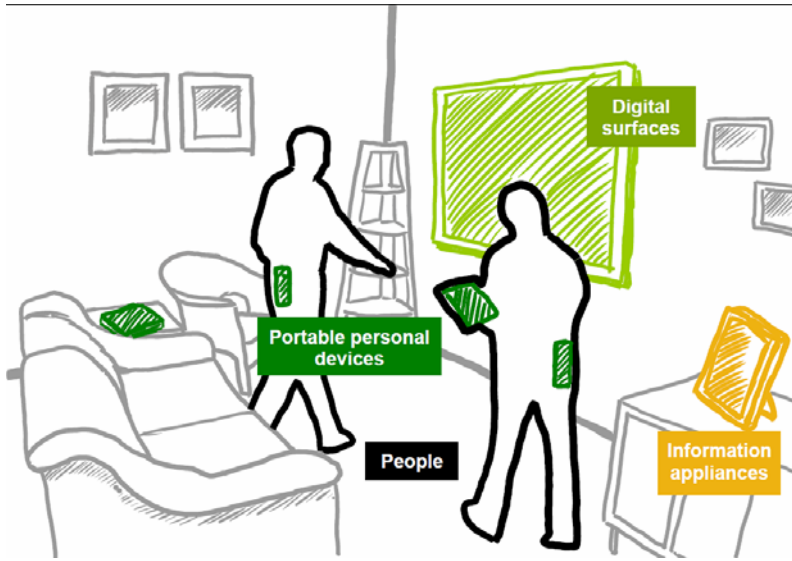


Fig. 1. Our scenario's ubicomp ecology, reproduced from [2]

Ambient to foreground interaction. The display catches John's attention, so he turns towards it and moves closer. As he does, the contents of the video player changes progressively. It shows increasingly more movies and more detail about each movie. When the display is within reach, touch controls appear so he can interact with it to choose and control videos [2, 9, 10, 16, 4].

Connection possibilities. John actually wants to play a video he has on his smart phone, so he pulls it out of his pocket. As he does, he sweeps it towards the wall display, but before it gets there, it is oriented towards the picture frame. The picture frame senses that the phone is oriented towards it, and a small icon appears on the picture frame that shows the two can connect. John keeps sweeping the phone across the room. When it is facing the large display, that display illustrates a connection icon.

Establishing connections. After a few moments of orienting the phone towards the display, the icon changes to show an image of his smart phone, indicating the two are connected [14, 15, 4]. Alternately, John could have just walked up to the display, pulled his phone out of his pocket, and touched the display with it [19].

Awareness of contents. As John approaches the wall display with his phone, the position of the icon on the wall display changes its position to follow the approaching phone. At the same time, the icon expands. First, the icon shows that it (i.e., the smart phone) contains several videos. As John moves even closer, the icon smoothly displays several images around it, each identifying one of the videos on the smart phone [2, 4].

Information transfer from up close. When he is within reach, John can transfer a video from his phone to the wall by several ways. He can just drag one off the icon, or he can pick and drop a video directly from his phone to the wall, or he can touch the relevant video image on the wall display with his phone [4, 3, 22].

Information transfer from afar. Alternately, John could have used a ray-casting technique as he was approaching the wall. He touches the desired video on his phone, and points the phone to the wall display: the video appears there as an icon as if it were ray-cast. A small flick of the phone transfers it across [4, 3].

Implicit actions. John turns back, sits on the couch and faces the screen. The system recognizes this, fills the screen with the last selected video, and starts playing it [16, 4, 3].

Attentive user interface. Part-way through the movie, John gets a phone call on his phone. He pulls it out of his pocket and presses it against his ear. The environment senses this relation of the phone to his head, and as a result the video play pauses. As soon as John puts the phone away and turns towards the screen, play resumes [11, 4, 3].

Multiple people. While John is watching, Shelly enters. Knowing that John is still watching but that Shelly had just entered, the wall display adjusts the video slightly to show its title. As Shelly moves closer, the running video is made a bit smaller and descriptive text appears near where Shelly is standing to explain what the movie is about [2, 10, 4].

Privacy. Shelly brings out her smart phone and, as with John, an icon appears on the display that shows it can connect. However, the contents are revealed as abstract entities. It is only when Shelly is close to the display that the actual contents are revealed in a way that is too small for John to see from afar, and where Shelly's body is shielding it.

Security. Sammy, a neighborhood kid, is visiting John's son. He knows John's system reacts to proximity, and would love to illicitly download some of John's action movies onto his smart phone, as his mom doesn't normally let him watch them. He goes into the room next door, where he stands on the other side of the wall holding the large display, brings out his phone, and points to it. However, the wall display ignores this, as the phone – even though it is close by – is oriented towards its rear.

Enhancing social security. Sammy then goes into the room and asks John if he could download some action movies. John, knowing about his mom's stance, says no but suggests he can grab some family movies. Sammy transfers one to his phone in the manner described above. Sometimes later, John is done with his movie and leaves. Sammy enters the room and tries to download an action movie. In this case, the display doesn't allow video transfers because John – who lives in the house – is not present.

The dark side #1. John re-enters later with his lover. They have no plans to watch video, but the screen turns on showing the latest action movies as they walk by. Distracted by the display, the intimate mode is broken.

The dark side #2. John is watching the hockey game (one of the option he can choose). His friend, who is on the road, had asked John to phone him and give him a play by play description of the game's progress. John wants to do this, but he cannot as the play pauses whenever he brings the phone to his ear.

The dark side #3. John has actually bought his large display at a highly reduced rate. Advertisers subsidized the cost, and John has agreed to have advertisements directed to him. Whenever John walks by, a commercial targeted to the things he usually buys appears. If he doesn't pay attention to it, the volume increases and the commercial gets busier, trying to attract his attention. It only goes away if he spends at least 10 seconds facing the ad [6].

5 Summary

The above examples are just a few of the opportunities (and annoyances) possible. Others exist in the current literature, and many more remain to be discovered. My hope is to stimulate your imagination, where you use these and other examples and concepts as a starting point to your own investigations.

Bio. Saul Greenberg is a Full Professor in the Department of Computer Science at the University of Calgary. He holds the NSERC/iCORE/Smart Technologies Industrial Chair in Interactive Technologies, and a University Professorship – a distinguished University of Calgary award recognizing research excellence. He received the CHCCS Achievement award in May 2007, and was also elected to the prestigious ACM CHI Academy in April 2005 for his overall contributions to the field of Human Computer Interaction.

While he is a computer scientist by training, the work by Saul Greenberg and his talented students typify the cross-discipline aspects of Human Computer Interaction, Computer Supported Cooperative Work, and Ubiquitous Computing. His many research contributions are bound by the common thread of situated interaction, which considers how computer technology fits within the fabric of people's day to day activities. This includes how such technology blends naturally in the flow of people's work practices, how people socialize and work together through technology, and how that technology fits within people's physical environment. He and his crew are well known for various significant contributions to the field, all necessary to pursue the broad goal of situated interaction:

- Articulation of design-oriented social science theories that serve as a requirements specification.
- Innovative and seminal system designs based on observations of social phenomenon.
- Toolkits enabling rapid prototyping of innovative groupware and ubiquitous appliances, and that exploit capabilities of multi-touch surfaces.
- Refinement of evaluation methods.

Acknowledgments. This research was partially funded by the iCORE/NSERC/SMART Technologies Industrial Research Chair in Interactive Technologies, and by the SURFNET NSERC Research Network. I am indebted to Nicolai Marquardt, Till Ballendat, Rob Diaz Marino, and Miaosen Wang: all contributed heavily to the ideas presented here.

References

1. Weiser, M.: Ubiquitous Computing. *Computer* 26, 71–72 (1993)
2. Greenberg, S., Marquardt, N., Ballendat, T., Diaz-Marino, R., Wang, M.: Proxemic interactions: the new ubicomp? *Interactions* 18, 42–50 (2011)
3. Ballendat, T.: Visualization of and Interaction with Digital Devices around Large Surfaces as a Function of Proximity. Diplom Thesis, University of Munich, Germany (2011)
4. Ballendat, T., Marquardt, N., Greenberg, S.: Proxemic interaction: designing for a proximity and orientation-aware environment. In: *ACM International Conference on Interactive Tabletops and Surfaces*, pp. 121–130. ACM Press, Saarbrücken (2010)
5. Marquardt, N.: Proxemic interactions in ubiquitous computing ecologies. In: *Proceedings of the Extended Abstracts of the 2011 Annual Conference on Human Factors in Computing Systems*, pp. 1033–1036. ACM Press, New York (2011)
6. Marquardt, N., Diaz-Marino, R., Boring, S., Greenberg, S.: The Proximity Tool-kit: Prototyping Proxemic Interactions in Ubiquitous Computing Ecologies. In: *Proc. ACM Conference on User Interface Software and Technology, UST 2011*, ACM Press, New York (2011)
7. Hall, E.T.: *The Hidden Dimension*. Anchor Books, New York (1963)
8. Reeves, B., Nass, C.: *The Media Equation: How People Treat Computers, Tele-vision, and New Media Like Real People and Places*. Cambridge University Press, Cambridge (1996)
9. Carsten, T.P., Röcker, C., Streitz, N., Stenzel, R., Magerkurth, C.: Hello.Wall – Beyond Ambient Displays. In: *Adjunct Proceedings of Ubicomp*, pp. 277–278 (2003)
10. Vogel, D., Balakrishnan, R.: Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. In: *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology*, pp. 137–146. ACM Press, USA (2004)
11. Shell, J.S., Selker, T., Vertegaal, R.: Interacting with groups of computers. *Commun. ACM* 46, 40–46 (2003)
12. Cooperstock, J.R., Smith, K.C.: Reactive Environments: Throwing away your keyboard and mouse. *Communications of the Association of Computing Machinery (CACM)* 40, 65–73 (1997)
13. Fitzmaurice, G.W.: Situated information spaces and spatially aware palmtop computers. *Commun. ACM* 36, 39–49 (1993)
14. Gellersen, H., Fischer, C., Guinard, D., Gostner, R., Kortuem, G., Kray, C., Rukzio, E., Streng, S.: Supporting device discovery and spontaneous interaction with spatial references. *Personal Ubiquitous Comput.* 13, 255–264 (2009)
15. Holmquist, L.E., Mattern, F., Schiele, B., Alahuhta, P., Beigl, M., Gellersen, H.-W.: Smart-Its Friends: A Technique for Users to Easily Establish Connections between Smart Artefacts. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) *UbiComp 2001*. LNCS, vol. 2201, pp. 116–122. Springer, Heidelberg (2001)

16. Ju, W., Lee, B.A., Klemmer, S.R.: Range: exploring implicit interaction through electronic whiteboard design. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, pp. 17–26. ACM Press, USA (2008)
17. Rekimoto, J., Ayatsuka, Y., Kohno, M., Oba, H.: Proximal Interactions: A Direct Manipulation Technique for Wireless Networking. In: Interact (2003)
18. Shoemaker, G., Tsukitani, T., Kitamura, Y., Booth, K.S.: Body-centric interaction techniques for very large wall displays. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, pp. 463–472. ACM Press, USA (2010)
19. Tandler, P., Prante, T., Müller-Tomfelde, C., Streitz, N., Steinmetz, R.: ConnecTables: Dynamic Coupling of Displays for the Flexible Creation of Shared Workspaces. In: Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology (UIST 2001), pp. 11–20. ACM Press, New York (2001)
20. Greenberg, S.: Toolkits and interface creativity. *Multimedia Tools Appl.* 32, 139–159 (2007)
21. Diaz-Marino, R., Greenberg, S.: The proximity toolkit and ViconFace: the video. In: Proceedings of the Extended Abstracts of the 28th International Conference on Human Factors in Computing Systems, pp. 4793–4798. ACM Press, USA (2010)
22. Rekimoto, J.: Pick-and-drop: a direct manipulation technique for multiple computer environments. In: Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology, pp. 31–39. ACM Press, USA (1997)

Voice Games: Investigation Into the Use of Non-speech Voice Input for Making Computer Games More Accessible

Susumu Harada^{1,*}, Jacob O. Wobbrock², and James A. Landay³

¹ IBM Research – Tokyo, 1623-14 Shimotsuruma, Yamato-shi,
Kanagawa-ken, 242-8502, Japan

² The Information School, DUB Group, University of Washington, Seattle,
Washington, 98195, USA

³ Computer Science and Engineering, DUB Group, University of Washington, Seattle,
Washington, 98195, USA

haradas@jp.ibm.com, wobbrock@u.washington.edu,
landay@cs.washington.edu

Abstract. We conducted a quantitative experiment to determine the performance characteristics of non-speech vocalization for discrete input generation in comparison to existing speech and keyboard input methods. The results from the study validated our hypothesis that non-speech voice input can offer significantly faster discrete input compared to a speech-based input method by as much as 50%. Based on this and other promising results from the study, we built a prototype system called the *Voice Game Controller* that augments traditional speech-based input methods with non-speech voice input methods to make computer games originally designed for the keyboard and mouse playable using voice only. Our preliminary evaluation of the prototype indicates that the Voice Game Controller greatly expands the scope of computer games that can be played hands-free using just voice, to include games that were difficult or impractical to play using previous speech-based methods.

Keywords: Computer games, accessible games, speech recognition, non-speech vocalization.

1 Introduction

Computer games today offer much more than just entertainment. They can provide social as well as educational value, and new genres of games such as “games with a purpose” (GWAP) [1] are emerging that combine gaming with productive output. However, almost all of these games are designed to be played using some form of manual input device, typically a keyboard for generating discrete input and a pointing device such as a mouse for generating continuous input. Consequently, people with

* The entirety of the work presented in this paper was conducted while the first author was affiliated with the Computer Science and Engineering department, DUB Group, University of Washington, Seattle, Washington, 98195, USA.

various forms of motor impairments who find it difficult or impossible to use manual input devices are excluded from reaping the benefits of such games. Is there a way to make these games accessible to people with motor impairments?

There are a number of compelling reasons to explore how to make computer games more accessible. First, hands-free access to computer games can benefit a wide range of people, including nearly 300,000 children under the age of 18 in the U.S. alone (or 1 in 250 children) who have been diagnosed with arthritis or other rheumatologic conditions [2]. Also, for people with more severe forms of motor impairment, computer games and game-like social applications (such as SecondLife [3]) may be one of the few viable avenues for entertainment and social interaction, both of which play an important role in improving quality of life [4]. Second, research into effective hands-free game input can also spur future designs of novel multimodal games. Third, an increased understanding of how to effectively control computer games hands-free can also inform how to better design general user interfaces for interaction using non-manual input modalities.

Strengths and Limitations of Speech Input for Game Control. Speech input is one of the hands-free input modalities that holds potential for addressing the above challenges. Unlike eye trackers or head trackers, a speech recognition system does not require elaborate or expensive hardware, and recent operating systems—such as versions of Microsoft Windows since Vista—include speech recognition software for free. Furthermore, the human language enables virtually a limitless number of words and phrases to be uttered, each of which can be mapped to a discrete command to be executed on a computer. This can be beneficial for controlling a number of computer games that require numerous keystroke combinations to be executed. Also, the ability to utter natural language commands compared to the need to remember and recall arbitrary keystroke combinations can be especially beneficial for novice users and may lead to a boost in performance [5].

However, speech input also has a number of inherent limitations that are preventing it from being an effective hands-free game control modality. First, the time it takes a person to complete uttering a word or a phrase can be a significant factor in games that require sub-second timing. The processing time required to recognize the spoken utterance also adds to this delay. Related to this is the limitation on the maximum number of utterances that can be recognized within a short period of time. When the need arises for dynamically issuing multiple input events at a rapid and varying rate, the per-utterance delay imposes a limit on how many such utterances can be recognized within the short time span.

Another major limitation of speech-based input is its inability to specify smoothly varying input. Output from a speech recognizer is discrete in nature, as the output is not generated until a word or a phrase has been recognized. The result of a single recognition is a single event, with the recognized utterance as the parameter. This discrete nature makes it significantly challenging to specify continuous input such as smooth motion of a pointer through two-dimensional space.

The current limitations of speech-based game control methods can be summarized in the context of the space of input signals expected by computer games. As mentioned earlier, most existing computer games are designed to be played using devices that can provide some combination of discrete input signals—such as a

keyboard—and continuous input signals—such as a mouse or a joystick. However, they vary widely in the degree of temporal and spatial fidelity demanded from each type of input signals. For discrete input signals, they can range from simply requiring execution of a few non-time-critical commands, to demanding precise and rapid execution of multiple and possibly simultaneous commands. For continuous input signals, they can range from simply requiring non-time-critical selection of a point, to demanding fast and fluid steering. “Strategy games” such as card games and puzzle games typically do not require fast or precisely-timed inputs in either of these two dimensions. On the other hand, “skill-and-action” games such as most arcade-based games often require both discrete and continuous input signals to be executed rapidly and fluidly. Existing speech-based game control methods are currently limited to offering hands-free input signals within the subset of this input signal space defined by the lower ends of the discrete and continuous input signal dimensions.

Potential of Non-speech Voice Input for Game Control. Fortunately, our previous work and work by others have shown that such limitations of traditional speech-based input methods, especially their inability to provide fluid continuous input, can be addressed effectively by augmenting them with the use of non-speech voice input [6-10]. In particular, the Vocal Joystick engine [11] can continuously track non-speech vocal properties including volume, pitch, and vowel quality to enable fluid interface manipulations such as smooth mouse pointer control. A longitudinal study involving participants with and without motor impairments found that people can learn the directional vowel mapping after a few hours of training, and even approach the performance of physical joysticks for pointing tasks [12]. The Vocal Joystick engine has also been successfully integrated into applications such as a hands-free drawing program called VoiceDraw [7], demonstrating its expressivity and controllability in the context of continuous steering tasks. In a sense, these systems fill a part of the void left by current speech-based methods in the computer game input signal space mentioned above, namely the far end along the continuous input dimension.

We felt that the expressivity of the Vocal Joystick engine could also be harnessed to expand voice modality’s coverage of the game input signal space along the *discrete* input signal dimension as well, by enabling faster, more responsive discrete input than speech-based methods. We hypothesized that by using the Vocal Joystick engine to generate discrete instead of continuous input events in response to directional vowel vocalizations, we could achieve significantly faster input speeds as compared to using spoken commands. Such a method would also nicely complement the continuous input mode already offered by the Vocal Joystick engine, as the same directional vowel sound mapping can be used, maintaining consistency and likely enhancing learnability. The combination of these two methods, along with traditional speech-based input methods, could greatly expand the coverage of hands-free voice-based computer game input methods.

Towards this end, we conducted a quantitative experiment to determine the performance characteristics of non-speech vocalization for discrete input generation in contrast to existing speech and keyboard input methods using an application we built called the *Vocal D-Pad*. While it was expected that non-speech voice input will be faster than a speech-based input method, the results from the study showed that non-speech voice input is significantly faster, with indications that it may further approach

the performance of keyboard input with additional training. Based on these promising results from the study, we built a prototype system called the *Voice Game Controller*, which combines the rapid discrete input capability of the Vocal D-Pad, the fluid continuous input capability of the original Vocal Joystick application, and the natural language input capability of traditional speech-based input. Our preliminary evaluation of the prototype indicates that the Voice Game Controller greatly expands the scope of computer games that can be played hands-free using just voice, to include games that were difficult or impractical to play using previous speech-based methods.

2 Related Work

There is little prior work on general voice-driven input techniques for computer games. Most of what exists can be classified as either tools that map speech commands to keystroke combinations, or individual games specifically designed to use voice input in their own way within the game.

2.1 Current Speech-Based Game Controls

Speech input in games has primarily been used for communication among multiple players [13, 14] or for command-and-control-style applications [5]. There have also been some tools developed to enable a player to use their voice to control various aspects of computer games. Most common are tools that translate spoken commands into specific actions, usually in the form of keystroke combinations. Tools such as *Shoot* [15], *Voice Buddy* [16], and *VR Commander* [17] have been specifically marketed for voice control of games, and map spoken commands to a sequence of keystrokes. Such functionality can be quite useful in games such as flight simulators in which there are a large number of possible commands that can be executed at any time, often mapped to obscure key combinations. Other general “speech macro” tools such as *Vocola* [18] and *WSR Macros* [19] offer similar functionality, but do not include any game-specific functionality. Most of these tools act as keystroke generators and are often used in conjunction with a physical keyboard or a mouse as an augmentative input modality. One limitation of these tools is that they cannot be used on their own to play games that require continuous input from pointing devices such as the mouse or joystick.

Aside from using speech for dictation or command-and-control, today’s commercial speech recognizers also enable the execution of pointing tasks using speech. Features such as the *MouseGrid* and *SpeechCursor* [20] offer the ability to control the mouse pointer to a limited extent. Although these features do enable basic pointing and clicking tasks to be performed using speech, they take significantly longer than the mouse [20], and their discrete nature makes them unsuitable for any task that requires rapid pointing or smooth steering such as games.

2.2 Games Designed to Be Controlled by Voice

There have been a number of games designed specifically to be controlled using the player’s voice. Most of them use relatively simple vocal features such as pitch and volume. Commercial games such as *Karaoke Revolution* [21] and *Rock Band* [22]

track the pitch of the players' singing and generate a score based on how close they match the target pitch track. *PAH!* [23] and *shout n dodge* [24] are Flash-based side-scrolling shooter games that continuously map the volume of the microphone input to the vertical position of a spaceship. *PAH!* also enables the user to shoot by making the plosive sound "pah." *Racing Pitch* [25] uses the pitch of the audio input to control the speed of a slot car, with higher pitch vocalization causing the car to travel faster. Igarashi [8] presents a set of simple games that demonstrate the use of pitch changes and sound detection to control simple movements of the game character. His games involve making rising or falling pitch sounds to move a character up or down, varying the duration of the vocalization to control how far a character moves, and making plosive sounds to shoot. In *Humming Tetris*, Sporka *et al.* [26] map various pitch inflection "gestures" to various controls such as move left, move right, and rotate.

While these games were designed from the onset specifically to be controlled by voice, and while we cannot expect many of the existing or future computer games to be redesigned in such a way, they provide insight into the types of non-speech vocal properties that can be used to map to existing game inputs.

2.3 The Vocal Joystick Engine

As mentioned in the introduction, one of the promising developments in the area of voice-based input has been the Vocal Joystick engine [11], which not only tracks the continuous vocal parameters such as volume and pitch, but also vowel quality. The Vocal Joystick engine samples audio input from the microphone every ten milliseconds, and for every frame, if the incoming sound is a vowel sound, reports the recognition probability for each of the nine vowel classes shown in Figure 1. These probabilities are then aggregated based on the radial arrangement of the vowel classes to yield an overall continuous directional vector, which can be used to steer a mouse pointer. The Vocal Joystick engine library [11] provides an API through which

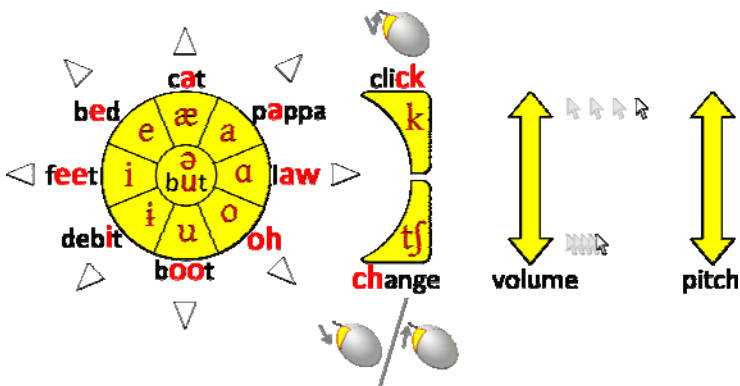


Fig. 1. The *Vocal Joystick* sound map showing the vocal features tracked by the Vocal Joystick engine. The *vowel compass* (left) shows the vowel sounds mapped to each radial direction. Red letters in each word approximate the corresponding sound. *Discrete sounds* such as “ck” and “ch” are also tracked, as well as changes in volume and pitch. The mapping to mouse pointer movements and events as used in the original *Vocal Joystick* application are shown.

third-party application developers can gain access to such information in real-time. The engine can also recognize non-vowel sounds, or discrete sounds, such as “ck” and “ch.” The original Vocal Joystick application enabled voice-driven control of the mouse pointer using the radial mapping of vowel sounds to directions as shown in Figure 1, as well as click and toggle of the left mouse button using the discrete sounds “ck” and “ch,” respectively. The volume was also mapped to the pointer speed, enabling the user to control both the pointer movement direction and speed fluidly and interactively. Studies have shown the Vocal Joystick to be effective for such basic mouse pointing and steering tasks [6, 12].

Due to the extremely short latency for vowel sound recognition compared to spoken word recognition, and the directional nature of the vowel sound mappings in the Vocal Joystick, we felt that non-speech voice-based *discrete* input can also offer significant advantages over speech-based methods. The following section describes the experiment we conducted to test this hypothesis.

3 Experimental Analysis of Non-speech Voice-Based Discrete Input

To better understand how non-speech voice input compares to keyboard and speech input as a method for generating discrete input signals, particularly under contexts requiring rapid and precisely-timed input, we conducted a reaction time (RT) experiment using an application we created called the *Vocal D-Pad*¹. The Vocal D-Pad is built using the Vocal Joystick engine library and generates emulated directional arrow-key events in response to recognized directional vowel sound utterances. The application can also emulate the corresponding directional keys being held down for the duration of the vocalization, as well as adjacent keys being held down simultaneously (e.g., the left and up keys being held down when the vowel sound for the upper-left direction is vocalized). The system also interfaces with the Microsoft Windows Speech Recognizer to enable mapping of spoken command input to emulated keyboard events, akin to the speech macro tools mentioned in the Related Work section.

The experiment involved the presentation of directional visual stimuli and user responses generated via one of three input modalities: *key* (four arrow keys on a keyboard), *speech* (words “up,” “down,” “left,” and “right”), and *voice* (directional vowel sounds based on the Vocal Joystick). The specific objectives and the setup of the experiment can be better understood in the context of the Model Human Processor [27], summarized next.

3.1 Experiment Background

Figure 2 shows a typical processing flow during a reaction time task in terms of the Model Human Processor. The Model Human Processor is an abstraction of the human cognitive-perceptual-motor system that can be useful in analyzing basic computer task performance. Upon the onset of a stimulus in a reaction time task, the user’s

¹ The name “D-Pad” comes from “directional pad,” a game input device consisting of four directional buttons.

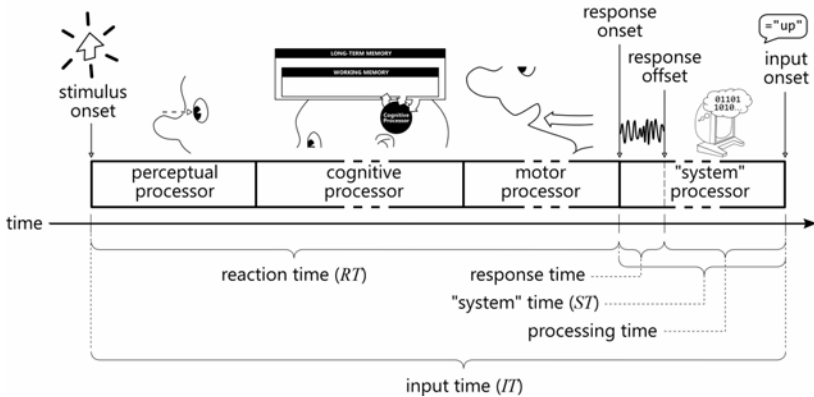


Fig. 2. Processing flow during a reaction time task based on the Model Human Processor [27] and the relevant measures. The *system processor* has been introduced to account for the time spent during response execution (*response time*) and recognition (*processing time*). At *input onset*, the recognition result is sent to the client application.

perceptual processor activates to process the relevant sensory input. The *cognitive processor* then deciphers the stimulus and determines the appropriate response. Finally, the *motor processor* converts the response action into motor signals and transmits them to the relevant motor systems such as the hand or the mouth. We also introduce a new module called the *system processor*, which is not in the original model since it is not part of the user. This represents any processing that the computer may need to perform on the input signal received from the user. The system processor does not add much time in cases such as button presses, but if the input is a spoken command, the system processor needs significantly more time to recognize what was uttered before it can provide its result as input to a client application. While the system processor time will be affected by the speed of the computer being used, the relative ranking of its value across different modalities should be unaffected.

Figure 2 also illustrates various measures that are of interest in a reaction time task. *Reaction time* is defined as the time between stimulus onset and *response onset*. Within the context of our experiment, response onset is defined as the earliest point when the system detects that some user response has been initiated (e.g., for a button-press response, when the button switch is closed, and for a voiced response, when the microphone first detects voicing activity). *System time* is the time between response onset and *input onset*—when the system finishes processing the response signal from an input device and generates a corresponding input signal for client applications. For keyboard input, the system time is a negligible value, but for speech input using a general recognizer, the value can be significant. We refer to the total elapsed time between stimulus onset and input onset as the *input time*.

There are two types of reaction time tasks that are commonly administered: *simple reaction tasks* and *choice reaction tasks*. In a simple reaction task, there is only one type of stimulus, and the user is instructed to issue a fixed response as soon as they perceive the stimulus. In a choice reaction task, the user is presented one of a number of possible stimuli, each with a corresponding response that must be correctly executed. The difference in the response times between these two tasks manifests itself in the cognitive processor time.

3.2 Hypotheses

Based on the model and representation presented above, we formulated the following hypotheses about the differences between keyboard, speech, and non-speech voice modalities with regards to their performance in reaction time tasks (summarized visually in Figure 3).

- *Hypothesis #1:* For trained users, the difference in reaction times across modalities is expected to be relatively small since the perceptual processor time and the cognitive processor time should be constant across modalities. In particular, reaction time for voice and speech modalities should be nearly identical given that the same motor system is involved and thus the same amount of time should be spent in the motor processor. The reaction time for the key modality may be slightly faster than that for voice or speech given that the vocal cord activation may take longer to generate an audible sound compared to the movement of a finger.
- *Hypothesis #2:* For a novice user under the choice reaction task using the voice modality, the unfamiliarity with the sound-to-direction mapping would most likely lead to a longer cognitive processor time and thus longer reaction time compared to the key or speech modality.
- *Hypothesis #3:* For both the simple and choice reaction tasks, the system processor time for the key modality is expected to be significantly faster than that for the speech modality, and the system processor time for the voice modality to lie somewhere in between, since non-speech vocalization should take less time to utter and be recognized compared to a speech command given its relative simplicity.

While we established the above three hypotheses in order to gain better insight into the nature of the differences among the three modalities' input times, we were also interested in the overall magnitude of those differences. We expected the key modality to be faster than the voice modality, and the voice modality to be faster than the speech modality, but the pragmatic question was, by how much, particularly in the context of computer games?

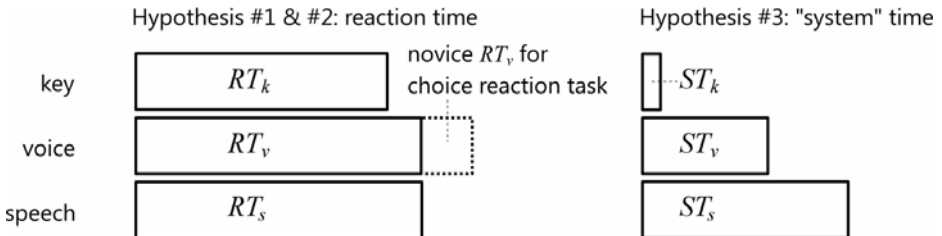


Fig. 3. Summary of the hypotheses of the experiment. Hypothesis #1 states that reaction times across modalities should be relatively similar, especially between voice and speech for trained users. Hypothesis #2 states that for a novice user, the voice modality may incur additional time to recall the unfamiliar sound-to-direction mapping. Hypothesis #3 states that the system time should be almost zero for the key modality, and that it should be significantly less for the voice modality compared to the speech modality.

3.3 Participants

Eight participants ranging in age from 21 to 34 were recruited to take part in the experiment. Two of the participants had motor impairments that affected the use of their hands for manipulating a keyboard and mouse (MI group), and the rest did not (NMI group). All eight participants had participated in prior Vocal Joystick user studies and had used the Vocal Joystick for an average of four hours each and had thus been familiarized with the directional vowel mappings. One of the MI participants (MI1) who had arthrogryposis multiplex congenita was unable to use the hands to manipulate a mouse or keyboard, but was able to use a mouth stick to press the keys on a keyboard. The other MI participant (MI2) had muscular dystrophy and was able to use the back of her fingers to press the keyboard keys.

3.4 Apparatus and Procedure

Each participant took part in a series of reaction time trials in which a test application presented them with a visual stimulus in the form of a blue arrow (100×100 pixels) on a computer screen (15" LCD monitor with resolution of 1400×900 pixels) pointing in one of the four cardinal directions. The participants were asked to respond to the stimulus in a manner dependent on a particular condition. Under all conditions, the test application waited for the correct key-down event and measured the elapsed time. The two independent factors and their levels were:

- Input modality: {key, voice, speech}
- Reaction task type : {simple, choice}

The *key* modality involved pressing one of the four arrow keys on a full-sized USB desktop keyboard. The *voice* modality involved uttering one of the four cardinal vowel sounds in the Vocal Joystick vowel compass (Figure 1), which the Vocal D-Pad processed to generate an emulated key-down event. Finally, the *speech* modality involved the utterance of a direction word (“up,” “down,” “left,” or “right”), also processed by the Vocal D-Pad in a similar fashion. The Windows Speech Recognizer’s general dictation grammar was disabled and only the grammar consisting of the four directional words was activated to minimize the error rate and maximize the processing speed. For both the voice and speech modalities, each participant underwent a basic adaptation process to tune the system to their voice. Under the voice modality, the user vocalized each of the four vowel sounds for two seconds while the Vocal Joystick engine recorded them and updated its acoustic model. Under the speech modality, the user read several paragraphs of passage as part of the Windows Speech Recognizer voice training wizard process.

Under the *simple* reaction task, the stimulus was always an up arrow, and the participant was instructed to respond as quickly as possible using a fixed response (up key for the *key* modality, the upward vowel sound for the *voice* modality, and the word “up” for the *speech* modality). The time between the end of one trial and the appearance of the next stimulus was randomized between one and two seconds (this range of pre-stimulus interval was found in past studies to yield the fastest reaction times while reducing predictability [28, 29]). Under the *choice* reaction task, the

participant was asked to generate a response matching the arrow stimulus as quickly and as accurately as possible. In this mode, the delay between trials was one second.

For all trials, both the reaction time and system time as defined in Figure 2 were measured. In the case of the key modality, the response onset and input onset were both considered to occur simultaneously when the key-down event was generated; therefore, its system time was always zero. In the case of both the voice and speech modality, the response onset was registered when the first voiced audio frame was detected by the Vocal Joystick engine, and input onset occurred when the corresponding emulated key-down event was registered.

3.5 Results

Aggregate results from the NMI group are presented first, followed by results from the MI participants. Figure 4 summarizes the results from the NMI group. The results are grouped first by the task type, with each bar corresponding to an input modality showing the reaction time and system time portions that made up the total input time. Henceforth, results reported as significant are at $p < .05$ level.

Differences in Reaction Times. For the simple reaction task, reaction times for voice and speech modalities were comparable as expected, since they both involve the same muscle groups. Reaction time for the key modality was significantly shorter as expected [30, 31]. This is most likely due to the vocal cord activation taking longer to generate an audible sound compared to the time it takes for the finger muscle activation to result in a key being pressed. This confirms our hypothesis #1. The average difference in these reaction times (176 milliseconds) can be interpreted as the approximate difference in the motor processor times between the manual and vocal modalities

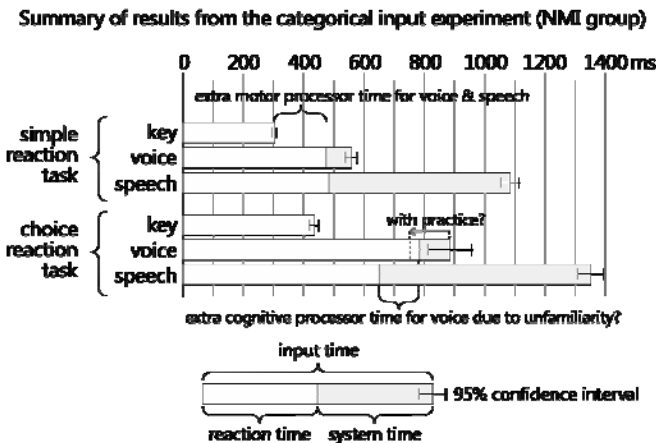


Fig. 4. Summary of the aggregate results from the categorical input experiment for the NMI group. The results are grouped by task type. Each bar shows the contribution from reaction time and system time towards the total input time, as defined in Figure 2. Error bars show 95% confidence intervals for total input times.

Under the choice reaction task, voice modality reaction time was slower than that for speech (782 milliseconds versus 649 milliseconds). This could be seen as supporting hypothesis #2, since even though the participants have had four hours of exposure to the Vocal Joystick, their association of the vowel sounds with their corresponding directions may not yet be automatic. The fact that this difference did not manifest itself in the simple reaction task indicates that the slowdown is most likely due to the extra cognitive processing required under the choice reaction task. Our prior work [12] has shown that users can fully memorize the directional vowel mappings in under 5 hours, but it may require more practice for recall of directional vowels to become as quick and as automatic as directional words. This suggests that with practice, voice modality reaction time under the choice reaction task could approach that of the speech modality. Figure 4 depicts this potential improvement on the choice reaction task input time bar.

Differences in System Times. The comparison of system times across task types within each modality verifies that the system processor time did not depend on the task type, as expected. The significantly shorter system time for the voice modality compared to the speech modality supports our hypothesis #3, and reflects the major advantage of the use of non-speech vocal input. While it has not been precisely determined how much of this difference is due to response time or processing time, it is most likely the case that the majority of it is accounted for by the difference in processing time.

Differences in Input Times. The comparison of the total input times reveals that overall, the key modality was the fastest (as expected), at 301 milliseconds for the simple reaction task and 433 milliseconds for the choice reaction task. However, the input time for the voice modality was also significantly faster than speech, by almost 50% in the simple reaction task and by 35% in the choice reaction task. The latter difference can be improved to 45% if the difference in cognitive processor time mentioned above is subtracted. While the voice modality input time still lags behind the key modality, that difference is significantly less than the improvement over the speech modality.

Results from MI Participants. The results from the two participants with motor impairments are summarized in Figure 5. For MI1, under the choice reaction task, reaction times across all three modalities were comparable (average of 1,050 ms after the learning effect adjustment described above). The fact that the key modality reaction time was comparable to voice and speech modalities is likely due to the difficulty of moving the mouth stick quickly in response to the stimulus. Also, under the choice reaction task, the voice modality input time was comparable to the key modality and almost 40% faster than the speech modality. Results for MI2 were similar as MI1, with voice modality input time under the choice reaction task (average of 780 ms after the learning effect adjustment) being only 25% slower than the key modality and almost 45% faster than the speech modality. These results offer a

promising sign for the comparative advantage of the voice modality over the key modality, especially for these participants with motor impairments.

Another interesting observation is the fact that the voice and speech input times for both MI1 and MI2 were longer than the average time for the NMI group. Our observations seem to suggest that this may be due to the effect of the MI participants' motor impairments on their lung capacity and their ability to rapidly vary vocal parameters. Further study is needed to determine the variability of such effect among individuals and how much of it can be accounted for via training.

Summary of Results. Figure 6 illustrates the results from the above study in a concrete example context of the game of Pac-Man. In the figure, the player's Pac-Man character (in yellow) is assumed to be at a standstill at a juncture, with the choice of moving up or down to move away from the chasing ghost character (in pink). The ghost character is approaching Pac-Man at a fixed speed, as indicated by the distance labeled "1 second." The multiple positions of the ghost character with corresponding input modality labels represent the closest point from Pac-Man that the ghost can be allowed to get before it is too late for the player to decide on the direction to move, execute the corresponding input, and have it be registered by the game in time

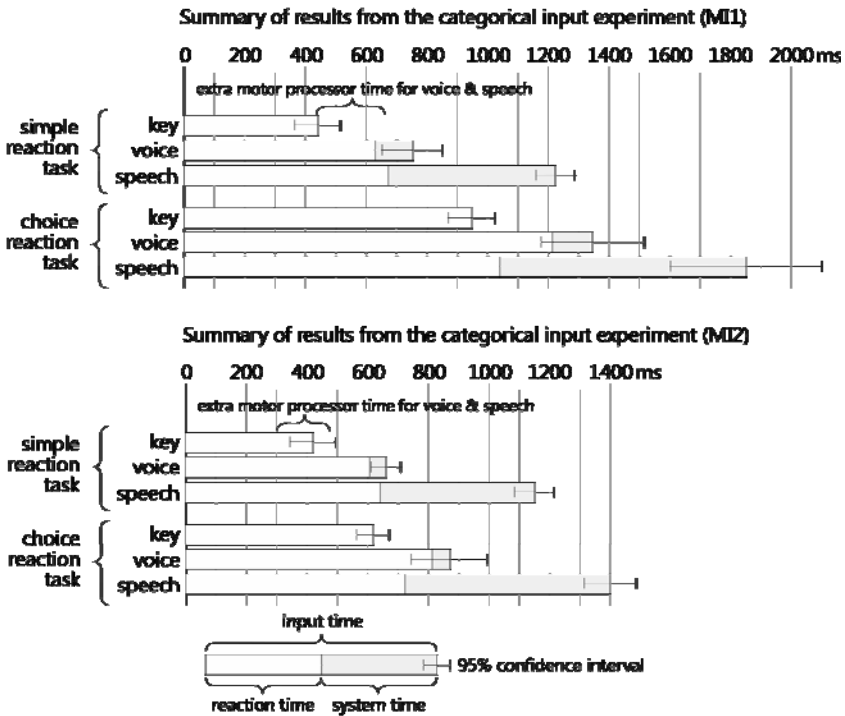


Fig. 5. Summary of the results from the categorical input experiment for the two participants in the MI group, presented in the same format as Figure 4.

to avoid contact with the ghost. In other words, these distances are the visual representations of the total input times under the choice reaction task shown in Figure 4. The position labeled “voice expert” shows the shortened distance that could be achieved if the reaction time for the voice modality is assumed to have become equal to that for the speech modality as described above. The figure highlights how significant an improvement the non-speech voice-based discrete input can offer over speech-based input, especially in the context of fast-paced time-critical computer games such as Pac-Man.

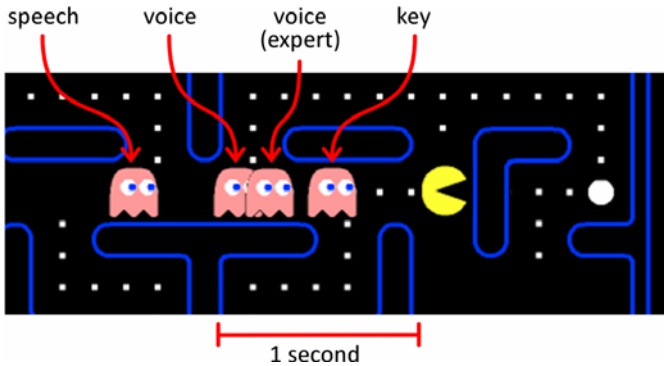


Fig. 6. An illustration of the implication of the result from the categorical input study in the context of a fast reaction time game Pac-Man. For each of the three input modalities, the figure shows the closest point that the “ghost” can be allowed to get to Pac-Man where there is just enough time for the player to decide on the direction of escape (in this case up or down), execute the corresponding response using the modality, and have the response be registered by the system.

4 The Voice Game Controller

The experimental analysis of non-speech voice-based discrete input yielded promising results over the speech-based input method. To leverage this potential and to situate it within the context of actual computer games, we built a system called the *Voice Game Controller* that extends the capability of traditional speech-based input by incorporating the rapid discrete input capability of the Vocal D-Pad and the fluid continuous input of the Vocal Joystick. The Voice Game Controller augments, rather than replaces, current speech-based input methods by retaining the option to use the dictation and command-and-control modes when non-time-critical speech input is appropriate.

The Voice Game Controller runs as a background process on the same computer as the one on which the game program is running. It translates the user’s vocalizations into keyboard, mouse and joystick (KMJ) signals that game programs expect as input. Figure 7 shows a high level representation of the Voice Game Controller architecture. The Voice Game Controller utilizes the Windows Speech Recognizer to recognize spoken command input, and the Vocal Joystick engine [11] to process non-speech vocalizations uttered by the user. Since each game requires different sets of input controls, the mapping between vocal input and corresponding KMJ signals is speci-

fied separately for each game in *Voice Game mapping profiles*. The *input mapping engine* passes the recognized voice input through a Voice Game mapping profile corresponding to the currently active game program, and generates the corresponding KMJ signal via the keyboard, mouse, and joystick emulator modules. Because the emulated KMJ signals are generated at the system level, they are application-independent and thus can be used to supply voice-driven KMJ input to non-gaming applications as well. From the game program’s perspective, it is as if the user is using a standard keyboard, mouse, or joystick.

The keyboard emulator module can generate key-down and key-up events for any key or set of keys, as well as repeated key-down events to emulate the user holding down a key. The mouse and joystick emulators can generate two-dimensional directional vectors (when in relative mode) or positional vectors (when in absolute mode), as well as button down and button up events corresponding to physical buttons on these devices. All these types of signals can be specified as the output of a mapping in a Voice Game mapping profile.

There are several ways in which spoken command input and non-speech vocal input can be mapped to the KMJ signals just described. First, spoken commands and non-speech discrete sound vocalization can be mapped to any of the key events for the keyboard emulator or to the button events for the mouse and joystick emulators. Second, non-speech vowel vocalization can be mapped to two-dimensional vectors based on the Vocal Joystick vowel compass mapping shown in Figure 1, with its magnitude possibly determined by the volume or pitch. This capability alone expands the realm of games playable by voice to include the numerous mouse-driven games. Third, non-speech vowel vocalization can also be mapped to key and button events similar to spoken commands and discrete sounds. This is done by determining the

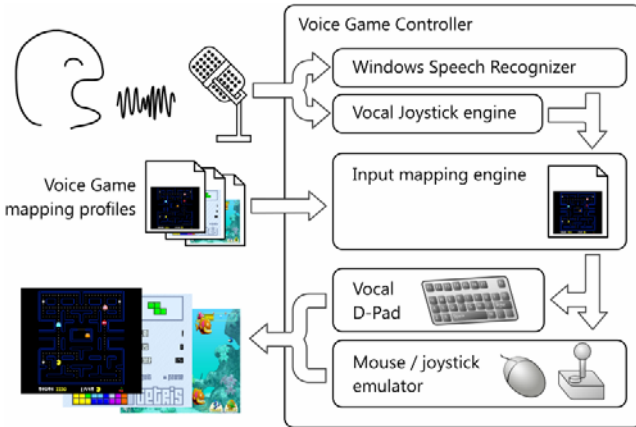


Fig. 7. Voice Game Controller architecture. User’s utterances are first processed by the Windows Speech Recognizer and the Vocal Joystick engine, and the corresponding keyboard/mouse/joystick signals are emulated based on the currently active *Voice Game mapping profile*.

vowel sound with the maximum probability every frame, instead of aggregating the probabilities across all vowels, and generating a binary output of whether or not that probability exceeds some threshold. If it does, it is treated as the *vowel onset* and a corresponding key or button event can be generated. The *vowel offset* event occurs for the current vowel when the vowel probability drops below the threshold, a different vowel's probability becomes higher, or the user stops vocalizing. The directional mapping of the Vocal Joystick vowel sounds make them ideal for mapping to the arrow keys in the context of games.

In the following section, we present our initial findings from a preliminary evaluation of the Voice Game Controller we conducted in order to assess its effectiveness in the context of four different games, comparing its performance to speech-driven control as well as to the participants' preferred input devices.

4.1 Preliminary Evaluation of the Voice Game Controller

Results from the comparative analysis of categorical input above showed that the voice input modality has the potential to offer significant performance gains over general speech input, approaching the performance of key input. To determine how this advantage translates to actual game play, we conducted a preliminary evaluation of the Voice Game Controller with the following four games to assess the viability of using the system to play real computer games.

Selected Games. We chose four games that lie at various points in the input signal space. Two Rooms [32] is a puzzle-based maze game in which the player controls square pieces using directional arrow keys. Although the time to complete each puzzle is considered, once the sequence is known, the emphasis is on how consistently and reliably the desired directional input can be generated. Pac-Man is a classic arcade game in which the player controls the Pac-Man character through a maze using directional arrow keys, attempting to eat all the white dots while avoiding four ghost characters. The game requires precise timing to avoid contact with a ghost character or to make a turn. The player's score corresponds to the number of white dots and frightened ghosts Pac-Man consumes. Tetris is a puzzle game in which the player rotates and positions falling two-dimensional tiles using arrow keys such that the tiles form solid horizontal layers, which then disappear from the playing field, yielding points for the player. It is necessary to be able to execute a rapid sequence of left or right movements to quickly move a piece into the desired position. Finally, Fish Tales [33] is a game in which the player controls a red fish in a two dimensional scene by using the mouse. The objective of the game is to eat fish that are smaller than the player's fish, while avoiding contact with larger fish. The score is based on the number of fish eaten. All of the fish move at varying speeds and directions, requiring the player to smoothly and rapidly steer the mouse pointer.

Evaluation Setup. Five of the eight participants who took part in the reaction time study, including the two participants with motor impairments (MI1 and MI2),

participated in the evaluation of the Voice Game Controller. Each participant came in for four 90-minute sessions separated by 48 hours or less. Throughout the sessions, we observed the users play each of the four games using three input modalities: non-speech voice input mode, speech-only input mode, and the user’s preferred input device. The speech-only input mode represents existing speech-based control methods, namely uttering directional words to emulate corresponding arrow key events and using features such as Speech Cursor and Mouse Grid for pointer control. For the preferred input device, the three participants in the NMI group chose the mouse and keyboard, while MI1 with arthrogryposis chose a mouth stick with the MouseKeys feature for controlling the pointer, and MI2 with muscular dystrophy chose the touchpad and keyboard.

During the first session, participants were shown how to use the Voice Game Controller and the speech-only input mode, and then introduced to the four games. For the remainder of the sessions, they played each of the games for 15 minutes at a time, spending 5 minutes on each of the three input modalities per game. At the end of the final session, a timed evaluation was conducted in which the participants played all games for five minutes each using each of the three input modalities. The speech input modality was omitted for the mouse-based Fish Tales game because the existing speech-based cursor control methods of Speech Cursor and Mouse Grid were too slow to be able to play the game at all. For the Two Rooms game, the measure of performance was based on the average completion time across all completed stages. The other three games were measured based on the average score achieved.

Results and Observations. Figure 8 summarizes the results from the evaluation described above. The data is presented in two graphs, with the graph on the right showing how the Voice Game Controller did compared to the speech modality, and the graph on the left showing how it did compared to the user’s preferred device. The horizontal axes represent how much better or worse the Voice Game Controller did with respect to the corresponding modality, with $\pm 0\%$ representing no difference in performance, and positive values indicating that it did better by that much percentage points.

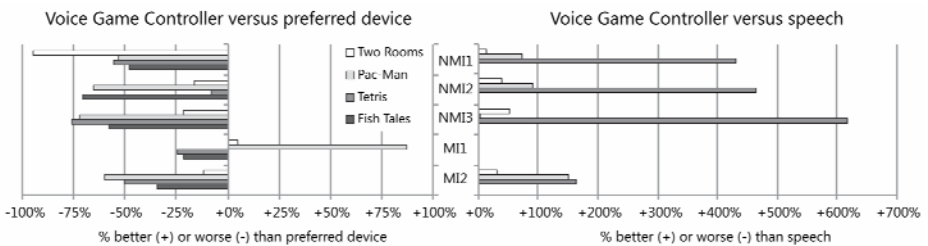


Fig. 8. Results from the Voice Game Controller user evaluation. The graph on the left shows how each participant scored in each game using the Voice Game Controller relative to their preferred device. The graph on the right shows the relative performance of the Voice Game Controller against speech-based input. $\pm 0\%$ indicates that the Voice Game Controller score was on par with the corresponding modality, and positive percentage indicates that the Voice Game Controller score was that much higher.

Compared to speech, the Voice Game Controller performed significantly better across all games for all participants (MI1's speech data was omitted due to incomplete tasks). For Fish Tales, the fact that the participants were able to play it at all using voice was a clear win over speech, with which the game could not be played. Within the NMI group, Voice Game Controller was 40% faster than speech in Two Rooms, and the average score was 50% higher in Pac-Man and 500% higher in Tetris. Within the MI group, Voice Game Controller was 32% faster than speech in Two Rooms, and the average score was 150% higher in Pac-Man and 160% higher in Tetris.

The better performance by the preferred input devices was expected, as was seen in previous comparative studies [20]. Within the NMI group, the Voice Game Controller performance was 40% worse than the preferred device for Two Rooms, and between 60% and 65% worse for the other three games. For the MI group, the Voice Game Controller performance was comparable to their preferred devices for Two Rooms, and between 60% and 70% for the other three games. While the preferred device performed better than the Voice Game Controller even for the MI group, both of the MI participants expressed that the fatigue induced by attempting to manipulate their physical devices was less desirable than the hands-free nature of the Voice Game Controller. None of the participants raised vocal fatigue as an issue during the study.

5 Conclusion

This paper presented results from our investigation into the viability of using non-speech voice-driven input for expanding the scope of computer games that can be controlled hands-free using voice only. Particularly, our experimental analysis into the use of non-speech vocalization for discrete input revealed a significant performance improvement over existing speech-based input method. Based on these findings, we built a prototype system called the Voice Game Controller that integrates the expressivity of traditional speech-based input with the fluid continuous input offered by the Vocal Joystick engine, as well as the non-speech voice-driven discrete input functionality of the Vocal D-Pad. The better performance obtained by the Voice Game Controller compared to standard speech-driven input in actual games suggests that voice input can become a viable modality for people with motor disabilities to play many of the mouse and keyboard-centric games that have previously been beyond reach for them. The expressiveness and hands-free nature of voice-driven input may also serve as a new game input modality for the general population, expanding the realm of interactive entertainment.

Acknowledgments. We would like to thank all the study participants who volunteered their time to try out our system. This work was supported in part by the National Science Foundation under grant IIS-0326382. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work was also supported by the University of Washington Royalty Research Fund.

References

1. Ahn, L.V., Dabbish, L.: Designing games with a purpose. *Commun. ACM*. 51, 58–67 (2008)
2. Sacks, J.J., Helmick, C.G., Luo, Y., Ilowite, N.T., Bowyer, S.: Prevalence of and annual ambulatory health care visits for pediatric arthritis and other rheumatologic conditions in the United States in 2001-2004. *Arthritis Rheum*. 57, 1439–1445 (2007)
3. Second Life Official Site, <http://secondlife.com/>
4. Riemer-Reiss, M.L., Wacker, R.R.: Factors associated with assistive technology discontinuance among individuals with disabilities. *Journal of Rehabilitation* 66, 44–50 (2000)
5. Tse, E., Greenberg, S., Shen, C., Forlines, C.: Multimodal multiplayer tabletop gaming. *Computers in Entertainment (CIE)* 5, 12 (2007)
6. Harada, S., Landay, J.A., Malkin, J., Li, X., Bilmes, J.A.: The Vocal Joystick: evaluation of voice-based cursor control techniques for assistive technology. *Disabil. Rehabil.: Assistive Technology* 3, 22 (2008)
7. Harada, S., Wobbrock, J.O., Landay, J.A.: VoiceDraw: a hands-free voice-driven drawing application for people with motor impairments. In: *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 27–34. ACM, Tempe (2007)
8. Igarashi, T., Hughes, J.F.: Voice as sound: using non-verbal voice input for interactive control. In: *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*, pp. 155–156. ACM, Orlando (2001)
9. de Mauro, C., Gori, M., Maggini, M., Martinelli, E.: Easy access to graphical interfaces by Voice Mouse. Università di Siena (2001)
10. Sporka, A.J., Kurniawan, S.H., Slavik, P.: Whistling User Interface (U3I). In: Stary, C., Stephanidis, C. (eds.) *UI4ALL 2004*. LNCS, vol. 3196, pp. 472–478. Springer, Heidelberg (2004)
11. Malkin, J., Li, X., Harada, S., Landay, J., Bilmes, J.: The Vocal Joystick Engine v1.0. *Computer Speech & Language* 25, 535–555 (2011)
12. Harada, S., Wobbrock, J.O., Malkin, J., Bilmes, J.A., Landay, J.A.: Longitudinal study of people learning to use continuous voice-based cursor control. In: *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, pp. 347–356. ACM, Boston (2009)
13. Gibbs, M., Wadley, G., Benda, P.: Proximity-based chat in a first person shooter: using a novel voice communication system for online play. In: *Proceedings of the 3rd Australasian Conference on Interactive Entertainment*, pp. 96–102. Murdoch University, Perth (2006)
14. Wadley, G., Gibbs, M.R., Benda, P.: Towards a framework for designing speech-based player interaction in multiplayer online games. In: *Proceedings of the 2nd Australasian Conference on Interactive Entertainment*, pp. 223–226. Creativity & Cognition Studios Press, Sydney (2005)
15. Shoot 1.6.4, <http://clans.gameclubcentral.com/shoot/>
16. Voice Buddy Interactive Voice Control Version 3.0, <http://www.edimensional.com/index.php?cPath=23>
17. VR Commander - Voice Command and Control for you Computer, <http://www.vrcommander.com/>
18. Vocola - A Voice Command Language, <http://vocola.net/>
19. Windows Speech Recognition Macros, <http://code.msdn.microsoft.com/wsrmacros>

20. Harada, S., Landay, J.A., Malkin, J., Li, X., Bilmes, J.A.: The Vocal Joystick: evaluation of voice-based cursor control techniques. In: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 197–204. ACM, Portland (2006)
21. Konami Digital Entertainment, Inc.: Karaoke Revolution, <http://www.konami.com/kr/>
22. Rock Band®, <http://www.rockband.com/>
23. PAH! - The first voice-controlled and activated game on the web, <http://games.designer.co.il/pah/>
24. shout n dodge, <http://www.weebles-stuff.com/games/shout+n+dodge/>
25. Racing Pitch, <http://jet.ro/games/racing-pitch/>
26. Sporka, A.J., Kurniawan, S.H., Mahmud, M., Slavík, P.: Non-speech input and speech recognition for real-time control of computer games. In: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 213–220. ACM, Portland (2006)
27. Card, S.K., Newell, A., Moran, T.P.: The Psychology of Human-Computer Interaction. Laurence Erlbaum Associates Inc., Hillsdale (1983)
28. Izdebski, K.: Effects of prestimulus interval on phonation initiation reaction times. *Journal of Speech and Hearing Research* 23, 485–489 (1980)
29. Shipp, T., Izdebski, K., Morrissey, P.: Physiologic stages of vocal reaction time. *Journal of Speech and Hearing Research* 27, 173–178 (1984)
30. Nebes, R.D.: Vocal versus manual response as a determinant of age difference in simple reaction time. *Journal of Gerontology* 33, 884–889 (1978)
31. Venables, P.H., O’connor, N.: Reaction times to auditory and visual stimulation in schizophrenic and normal subjects. *Q. J. Exp. Psychol.* 11, 175 (1959)
32. Armor Games: Two Rooms, <http://armorgames.com/play/3006/two-rooms>
33. Fish Tales, <http://flashgamesite.com/play334game.html>

GraVVITAS: Generic Multi-touch Presentation of Accessible Graphics

Cagatay Goncu and Kim Marriott

Clayton School of Information Technology, Monash University
{cagatay.goncu, kim.marriott}@monash.edu.au

Abstract. Access to graphics and other two dimensional information is still severely limited for people who are blind. We present a new multimodal computer tool, GraVVITAS, for presenting accessible graphics. It uses a multi-touch display for tracking the position of the user's fingers augmented with haptic feedback for the fingers provided by small vibrating motors, and audio feedback for navigation and to provide non-geometric information about graphic elements. We believe GraVVITAS is the first practical, generic, low cost approach to providing refreshable accessible graphics. We have used a participatory design process with blind participants and a final evaluation of the tool shows that they can use it to understand a variety of graphics – tables, line graphs, and floorplans.

Keywords: graphics, accessibility, multi-touch, audio, speech, haptic.

1 Introduction

Graphics and other inherently two dimensional content are ubiquitous in written communication. They include images, diagrams, tables, maps, mathematics, plots and charts etc. They are widely used in popular media, in workplace communication and in educational material at all levels of schooling. However, if you are blind or suffer severe vision impairment your access to such graphics is severely limited. This constrains enjoyment of popular media including the web, restricts effective participation in the workplace and limits educational opportunities.

There are a number of different techniques for allowing people who are blind to access graphics, the most common being tactile graphics presented on swell or embossed paper. We review these in Section 3. However, it is fair to say that none of these are widely used and that currently there is no reasonably priced technology or tool which can be effectively used by someone who is blind to access graphics, tables and other two-dimensional content. This is in contrast to textual content, for which there exist computer applications widely used by the blind community. For instance, DAISY provides access to textbooks and other textual material using speech or refreshable Braille displays and Apple's VoiceOver screen reader provides accessible access to the text in webpages.

The main contribution of this paper is to present the design and evaluation of a new tool for computer mediated access to accessible graphics. The great advantages of our

tool are that it is relatively cheap to construct and costs virtually nothing to operate, provides a generic approach for presenting all kinds of 2-D content, can support dynamic, interactive use of graphics and could be integrated with existing applications such as DAISY.

GraVVITAS (for Graphics Viewer using Vibration, Interactive Touch, Audio and Speech) is a multi-modal presentation device. The core of *GraVVITAS* is a touch sensitive tablet PC. This tracks the position of the reader's fingers, allowing natural navigation like that with a tactile graphic. Haptic feedback is provided by small vibrating motors of the kind used in mobile phones which are attached to the fingers and controlled by the tablet PC. This allows the user to determine the position and geometric properties of graphic elements. The tool also provides audio feedback to help the user with navigation and to allow the user to query a graphic element in order to obtain non-geometric information about the element.

We have used a user-centered and participatory design methodology, collaborating with staff from Vision Australia¹ and other relevant organizations and blind participants at all stages in the design and development of the tool. We believe participatory design with blind participants is vital for any project of this kind since our experiences, and previous research suggest that people who have been blind from an early age may have quite different strategies for understanding graphics to people who are sighted [25]. The results of our evaluation of *GraVVITAS* are very positive: our blind participants learnt to use the tool to understand a variety of graphics including tables, line graphs and floorplans.

2 Design Requirements

In this section we detail our three initial design requirements. These were developed in collaboration with staff at Vision Australia. The first design requirement is that the computer tool *can be used effectively by people who are blind to read an accessible version of a wide range of graphics and 2D content*. This means that the accessible version of the graphic should contain the same information as the original visual representation. However simple information equivalence is quite a weak form of equivalence: a table and a bar chart presenting the same data are equivalent in this sense. We require a stronger form of equivalence in which the spatial and geometric nature of the original graphic is maintained, so that the blind viewer of the accessible version builds up an internal spatial representation of the graphic that is functionally equivalent to that of the sighted viewer. Such *functional equivalence* is important when graphics are being used collaboratively by a mixture of sighted and blind people, say in a class room or workplace, or when contextual text explains the graphic by referring to the graphic's layout or elements.

Functional equivalence also means that the accessible graphic is more likely to maintain at least some of the cognitive benefits that sighted readers obtain when using a graphic instead of text. Starting with Larkin and Simon [21] many researchers have investigated the differences between graphics and text and the benefits that can make

¹ Vision Australia is the primary organization representing people with vision impairment in Australia and a partner in this project.

graphics more effective than text [31, 33, 30]. Such benefits include: geometric and topological congruence, homomorphic representation, computational off-loading, indexing, mental animation, macro/micro view, analogue representation and graphical constraining. While it is unlikely that all of these benefits will be displayed by the accessible representation we believe that many will be [14].

The second design requirement is that *the tool is practical*. This means that it has to be inexpensive to buy and to operate, can be used in classrooms, home and work environments, and can be integrated with other applications such as screen readers.

The final design requirement is that *the tool supports interactive, active use of graphics*. This means that the tool must have a rapidly refreshable display so that it supports the kind of interactive use of graphics that sighted users now take for granted: interactive exploration of a graphic at different levels of detail; creation and editing of graphics; and dynamic presentation of graphics created by applications like graphing calculators or spreadsheet tools.

3 Background

We now review the main previous approaches to accessible graphics and evaluate them with respect to our three design requirements. As a first step it is useful to review different characteristics of the relevant human perceptual subsystems [6, 16].

The visual subsystem has sensors that receive light and provide visual information such as shape, size, colour, intensity and position. It needs no physical contact with objects to acquire this information. It has a wide area of perception that provides parallel information in a continuous flow and within this is a narrow area (the fovea) which can detect highly detailed information.

The haptic subsystem requires physical contact with objects to acquire information. Cutaneous sensors on the skin detect touch and temperature, while the kinesthetic sensors on the muscles and joints of the body sense motion. The haptic subsystem can provide much of the same information as the visual subsystem (shape, size, texture and position) and haptic input can lead to internal spatial representations that are functionally equivalent to those obtained from visual input [4].

The aural subsystem has sensors that receive aural information such as audio, and speech. It is more effective in acquiring sequential stimuli. Since the aural subsystem provides binaural hearing it can also locate the source of a stimulus. It does not need to have a physical contact with the objects to acquire this information.

Tactile graphics are probably the most frequently used approach to accessible graphics and are commonly used in the education sector. They allow the viewer to feel the graphic and have been in use for over 200 years [10]. Tactile graphics are usually displayed on embossed tactile paper in which embossers punch the paper with varying height dots to create raised shapes or thermo-form (swell) paper which contains thermo capsules that rise when heat is applied. Both of these are non-refreshable media.

Much less commonly, tactile graphics can be displayed on electro-mechanical refreshable displays [36]. These have multiple lines of actuators that dynamically change in time. When the display is activated, the user traces the area to feel what is on the display. These refreshable displays are primarily designed for presenting Braille. Larger displays suitable for presenting tactile graphics are expensive (e.g. A4 size displays are around US \$20,000) and have quite low resolution.

One limitation of a pure tactile presentation is that text must be presented as Braille. This takes up considerable space and many blind users cannot read Braille. It can also be difficult to use easily distinguishable textures when translating a graphic that makes heavy use of patterns and colour. From our point of view, however, the main limitation of tactile graphics is that they are typically created on request by professional transcribers who have access to special purpose paper and printers. As a result they are expensive and time consuming to produce. For instance, transcription of the graphics in a typical mathematics textbook takes several months and is estimated to cost more than US \$100,000. Furthermore non-refreshable media do not support interactive use of graphics.

TGA [19] overcomes the need for professional transcribers by using image processing algorithms to generate tactile graphics. Text in the image is identified and replaced by the Braille text and the visual properties such as colours, shading, and textures are simplified. The image is then uniformly scaled to satisfy the required fixed size of the Braille characters. However, it still requires access to expensive special purpose paper and printers or a refreshable display. Furthermore, because of the large amount of scaling that may be required to ensure that the Braille text does not overlap with other elements the results are sometimes unsatisfying.

Touch sensitive computing devices like the IVEO [13] and Tactile Talking Tablet (TTT) [20] are a relatively new development. These allow a tactile graphic to be overlaid on top of a pressure-sensitive screen. When reading the user can press on an element in the tactile overlay to obtain audio feedback. The main advantage is that audio feedback can be used instead of Braille. However, the use of these devices is limited, requires expensive tactile overlays and does not support interactive use of the graphic.

To overcome the need for expensive tactile overlays some tools have been developed that rely on navigation with a joystick or stylus. A disadvantage of such approaches is that unlike tactile graphics, they do not allow multi-hand exploration of the graphic since there is a single interaction point for navigation.

One of the most mature of these is TeDub (Technical Drawings Understanding for the Blind) [27]. It is designed to present node-link diagrams such as UML diagrams. TeDub uses an image processing system to classify and extract information from the original drawing and create an internal connected graph representation through which the user can navigate with a force feedback joystick by following links. Speech is used to describe the node's attributes. A key limitation from our point of view is that the navigation and interaction is specialized to node-link diagrams and is difficult to generalize to other kinds of graphics.

The VAR (Virtual Audio Reality) [12] tool also provides a joystick for navigation. It allows the user to perform tasks on a graphical user interface. The elements in the visual interface are represented by short audio representations placed in a 3D space. The user navigates in this 3D space using the joystick. During the tracing, audio associated to elements are played through the headphones. In MultiVis, which has a similar design, the authors used a force-feedback device and non-speech audio to construct and provide quick overviews of bar charts [23]. A key limitation of VAR and MultiVis is that they are specialized to a particular kind of application.

In another study, a tool using a graphics tablet and a VTPlayer tactile mouse is evaluated [37] for the presentation of bar charts. The user explored a virtual bar chart on a graphics tablet using a stylus. Based on the position of the stylus, the two tactile

arrays of Braille cells on the mouse, which was held in the other hand, were activated. The activation of the pins in these cells was determined by the pixel values pointed by the stylus. Speech audio feedback was also provided by clicking the button on the stylus. The tool had the advantage that it was inexpensive to buy and cheap to run. Although designed for bar charts it could be readily generalised to other graphics. However, we believe that because the interaction is indirect (through a mouse controlling a cursor that the user cannot see) it would be quite difficult to learn to use. Another limitation is that it provides only a single point of interaction.

In [22] a tool for navigating line graphs was presented. This used a single data glove with four vibrator motors. The motors were not used to provide direct haptic feedback about the graphic but rather were used to inform the user on which direction to move their hand in order to follow the line graph.

A hybrid tactile overlay/haptic approach was employed in a networked application that allowed blind people to play a board game called Reversi (also called Othello)[26]. This used a touch screen with a tactile overlay to present the board and dynamic haptic and audio feedback to present the position of the pieces on the board.

Layered audio description of the graphic and its content is a reasonably common technique for presentation of graphics to blind people. This is typically done by trained transcribers and so is expensive and time consuming. It also has the great disadvantage that functional equivalence is lost. Elzer et al [9] have developed an application for automatically generating an audio description of a bar chart summarizing its content. This overcomes the need for a trained transcriber. While clearly useful, for our purposes the disadvantages are that the application is specialized to a single kind of information graphic and that it does not preserve functional equivalence.

Thus we see that none of the current approaches to presentation of accessible graphics meet our three design requirements: there is a need for a better solution.

4 Design of GraVVITAS

We used a participatory design approach in collaboration with blind participants to design our tool. We initially planned to use a more formal usability testing approach but we found that we were often surprised by what our blind participants liked or disliked, and so found it difficult to foresee some of the problems in the interface. Therefore we instead used a participatory design process [18] in which the design evolved during the course of the usability study and was sometimes changed during the user evaluations because of participant feedback.

It is worth pointing out that all approaches to presenting accessible graphics, including tactile graphics, require the blind user to spend a considerable amount of time learning to use the approach. This is a significant difficulty when evaluating new tools since it is usually not practical to allow more than a few hours training before a participant uses the tool. We partially overcame this problem by using the same participants in multiple user studies meaning that they had more experience with the tool.

Since there are relatively few blind people and it is often hard for them to travel, it is quite difficult to find blind participants (also pointed out in [32, 28]). Hence the number of participants was necessarily quite small—between 6 and 8 for each usability study. Participants were recruited by advertising the study on two email lists for

print-disabled people in Australia, and we used all who responded. They were all legally blind and had experience reading tactile graphics. They were aged between 17 and 63. Participants were asked to sign a consent form which had previously been sent by email to them and which they were given a Braille version of on the day. This also provided a short explanation of the usability study and what type of information would be collected.

4.1 Basic Design

One of the most important design goals for GraVVITAS was that it should allow, as far as possible, the blind user to build a functionally equivalent internal spatial representation of the graphic. We have seen that a haptic presentation allows this [4]. Previous studies have shown that blind participants prefer tactile presentations to audio [15] and audio is preferred in exploration and navigation tasks. All of our participants felt that tactile graphics were the most effective way that they knew of for presenting graphics to the blind.

We believe that one reason for the effectiveness of tactile graphics is that they allow natural navigation and discovery of geometric relationships with both hands and allow the use of multiple fingers to feel the geometric attributes of the graphic elements. The use of both hands allows semi-parallel exploration of the graphic as well as the use of one hand as an anchor when exploring the graphic. Both of these strategies are common when reading Braille and tactile graphics [11, 8].

However as we have noted, tactile graphics or overlays are expensive to produce and are non-refreshable so they do not support interactive use of the graphic. What is required is a low-cost dynamic tactile display that supports exploration with multiple hands and fingers. Recent advances in touch screen and haptic feedback devices finally allow this.

Our starting point was a touch sensitive tablet PC which tracks the position of the reader's fingers. We used a Dell Latitude XT² which is equipped with NTrig DuoSense dual-mode digitizer³ which supports both pen and touch input using *capacitive* sensors. The drivers on the tablet PC allowed the device to detect and track up to four fingers on the touchscreen. We allowed the user to use the index and middle finger of both the left and right hand.

A key question was how to provide haptic feedback to the reader's fingers so that they could feel like they were touching objects on the touchscreen. In recent years there has been considerable research into haptic feedback devices to increase realism in virtual reality applications including gaming, and more recently to provide tactile feedback in touch screen applications [2]. The main approaches are electromechanical deformation of the touch screen surface, mechanical activation applied to the object (stylus or finger) touching the surface, and electro-vibration of the touch screen, e.g. see [1]. In the longer term (i.e. 2+ years) there is a good chance that touch screens will provide some sort of dynamic tactile feedback based on electromechanical deformation or electro-vibration. However, during the time we have been developing GraVVITAS, mechanical activation applied to the fingers touching the screen was the most mature and reliable technology for supporting multi-touch haptic feedback.

² <http://www.dell.com>

³ <http://www.n-trig.com>

We therefore chose to provide haptic feedback by using a kind of low cost data glove with vibrating actuators. To do so we attached small vibrating motors of the kind used in mobile phones to the fingers and controlled these from the tablet PC through an Arduino Diecimila board⁴ attached to the USB port. Since the touchscreen could track up to four fingers there were four separately controlled motors. The amount of vibration depended on the colour of the graphic element under the finger and if the finger was over empty space there was no vibration.

One difficulty was that when there are more than four fingers on the touch screen the device behaved inconsistently and fingers touching the touchscreen were not always detected. To shield unwanted fingers, we used a cotton glove. The tool is shown in Figure 1. Detection of fingers remained an issue for some users who needed to be trained to flatten their finger tips to be properly detected by the touchscreen. During the training session we suggested that users lift their fingers up and put them down again to reset the finger assignment if they suspected one of their fingers was not properly detected. This meant that it took some time for some participants to get used to the tool.

Probably the most technically challenging part of the implementation was determining in real-time which fingers were touching the tablet and which finger corresponded to which touchpoint on the device. Knowing this was necessary for us to provide the appropriate haptic feedback to each finger. We stored the maximum and average vector difference between the stroke sequences on the device. Based on these differences we used a Bayesian approach which chose the most probable feasible *finger configuration* where a finger configuration is a mapping from each stroke sequence to a particular finger. A configuration was infeasible if the mapping was physically impossible such as assigning the index and middle finger of the same hand to strokes that were sometimes more than 10cm apart. There was a prior probability for each finger to be touching the device and a probability of a particular finger configuration based on an expected vector difference between each possible pair of fingers. We also used the area of the touch points, and the angle between them in the calculations. The approach was quite effective.

One disadvantage of using a haptic presentation of a graphic is that because of the sequential movement of hands and fingers involved in perception, acquisition of in-

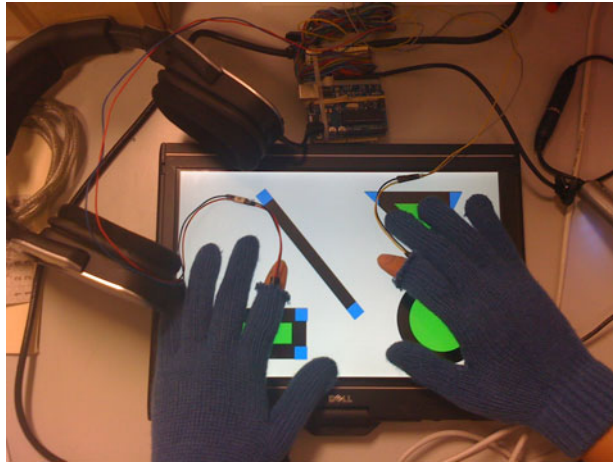


Fig. 1. Using GraVVITAS to view a diagram

⁴ <http://arduino.cc>

formation is slower and less parallel than vision. Also, because there is no haptic equivalent of peripheral vision, the position of previously encountered objects must be stored in memory [34]. To partially address this problem, we decided to provide audio feedback in order to help the user with navigation and to obtain an overview of the graphic and its layout. The use of audio means that the user can obtain an view without having to physically touch the elements.

Another disadvantage of a purely haptic presentation is that it is difficult to represent non-geometric properties of elements and text. While Braille can be used it takes up a lot of space and cannot be read by many users. To overcome this we decided to provide audio feedback when the viewer queries graphic elements on the display. This was similar to TTT or IVEO.

The tool displays graphic content specified in SVG (the W3C standard for Scalable Vector Graphics) on a canvas which is implemented using Microsoft Windows Presentation Framework. The canvas loads a SVG file and use the metadata associated with the shapes to control the tool behaviour. The metadata associated with a shape is: its ID, the vibration level for the edges and audio volume level for the interior of the shape and for its boundary, the text string to be read out when the shape is queried, and the name of a (non-speech) audio file for generating the sound associated with the shape during navigation. The SVG graphics could be constructed using any SVG editor: we used Inkscape⁵. The only extra step required was to add the metadata information to each shape. We did this using Inkscape's internal XML editor.

4.2 Haptic vs. Audio Feedback

In our first trials with the tool we experimented with the number of fingers that we attached the vibrating motors to. We tried: (a) only the right index finger, (b) the left and right index fingers, and (c) the left and right index and middle fingers. Our experience, corroborated by feedback from a single blind participant, was that it was beneficial to use fingers on both hands but that it was difficult to distinguish between vibration of the index and middle finger on the same hand. We first tried attaching the vibrating devices to the underside and then to the top of the finger but this made little difference. Our experience is that, with enough practice, one can distinguish between vibration on all four fingers but this takes many hours of use. We therefore decided to use the tool with two fingers—the left and right index fingers—as we would not be able to give the participants time to learn to use four fingers when evaluating the tool.

Given that we decided only to provide haptic feedback for the left and right index finger, a natural question to investigate was whether stereo audio feedback might be better. To determine this we implemented an audio feedback mode as an alternative to haptic feedback. This mode was restricted to the use of one finger or two fingers on different hands. In audio mode if the user touches an object on the screen then they will hear a sound from the headphones. If they use one finger they will hear a sound coming from both headphones while if they use two fingers then they will hear a sound on the left/right headphone if their left/right finger is on an element. The sounds associated with objects were short tones from different instruments played in a loop. They were generated using the JFugue library.

⁵ www.inkscape.org

We conducted a usability study to investigate whether audio or haptic feedback was better for determining the geometric properties (specifically position and shape) of graphic elements. The study used simple graphics containing one to three geometric shapes (line, triangle, rectangle and circle) such as those shown in Figures 2, and 3. Each shape had a low intensity interior colour and a thick black boundary around it. This meant that the intensity of the haptic or audio feedback was greater when the finger was on the boundary.

We presented the graphics to each participant in the two different modes—audio and haptic—in a counterbalanced design. For each mode the following two-step procedure was carried out. First we presented the participant with one training graphic that contained all of the different shapes. In this step we told them what shapes were on the screen and helped them to trace the boundaries by suggesting techniques for doing so and then letting them explore the graphic by themselves. Second, the participant was shown three graphics, one at a time and asked to explore the graphic and let us know when they were ready to answer the questions. They were then asked to answer two questions about the shapes in the graphic:

1. How many objects are there in the graphic?
2. What kind of geometric shape is each object?

The times taken to explore the graphic and then answer each question were recorded as well as their answers. After viewing and answering questions about the graphics presented with the audio and haptic interaction modes the participants were asked which they preferred and invited to give comments and explain the features that influenced their preference.

Eight participants completed the usability study. We found that 6 out of 8 participants preferred haptic feedback. Error rates with audio and haptic feedback were very similar but the time to answer the questions was generally faster with haptic feedback.

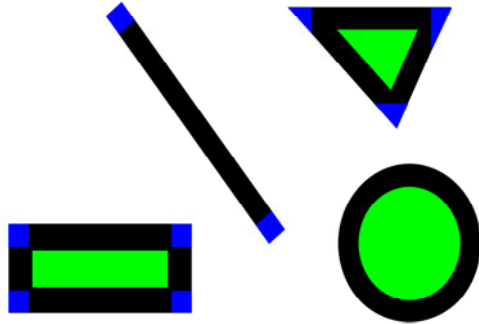


Fig. 2. Example graphic used in haptic vs audio feedback usability study

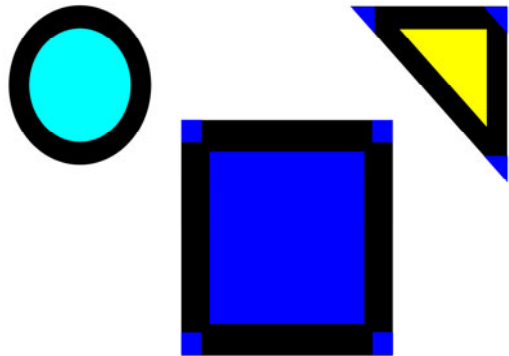


Fig. 3. Example graphic used in audio interface design usability study

These results need to be considered with some care because they were not statistically significant because of the small number of participants.

Another caveat is that we slightly modified the presentation midway through usability study. This was because the first three participants had difficulty identifying the geometric shapes. The reason was that they found it difficult to determine the position and number of vertices on the shape. To overcome this in subsequent experiments object vertices were given a different color so that the audio and haptic feedback when touching a vertex differed from that for the boundary and the interior of the shape. This reduced the error count to almost zero in the subsequent participants.

We observed that participants used two quite different strategies to identify shapes. The first strategy was to find the corners of the shapes, and then to carefully *trace* the boundary of the object using one or two fingers. This was the strategy we had expected.

The second strategy was to use a single finger to repeatedly perform a quick horizontal and/or vertical *scan* across the shape, moving the starting point of the finger between scans slightly in the converse direction to that of the scan. Scanning gives a different audio or haptic pattern for different shapes. For instance, when scanning a rectangle, the duration of a loud sound on an edge, a soft sound inside the shape, and another loud sound on the other edge are all equal as you move down the shape. In contrast for a triangle the duration of the soft sound will either increase or decrease as you scan down the shape. This strategy was quite effective and those participants who used it were faster than those using the boundary tracing strategy.

As a result of this usability study we decided to provide haptic feedback (through the vibrating motors) rather than audio feedback to indicate when the user was touching a graphic element. The choice was because of user preferences, the slight performance advantage for haptic feedback, because haptic feedback could be more readily generalized to more than two fingers, and because it allowed audio feedback to be used for other purposes.

4.3 Design of the Audio Interface

The next component of the interface that we designed was the audio interface. We investigated the use of audio for two purposes: to provide non-geometric information about a graphic element and to help in navigation.

The initial interface for obtaining non-geometric information about a graphic element was similar to that used in IVEO or TTT. If a finger was touching a graphic element the user could query the element by “twiddling” their finger in a quick tiny circular motion around the current location without lifting it up. This would trigger the audio (speech or non-speech) associated with the element in the SVG file. Audio feedback could be halted by lifting the finger from the tablet. Audio feedback was triggered by whichever finger the user twiddled and could come from more than one finger.

Designing the interface for determining the position of elements in the graphic using audio was more difficult and we developed two quite different techniques for doing this.

The first technique was to generate a *3D positional audio* based on the location of one of the fingers on the touchscreen. This use of 3D audio was based on initial

conversations and studies with blind people who said they liked the use of 3D audio in computer games [35]. When the user was not touching an element, they would hear through the headphones the sound associated with the graphic elements within a fixed radius of the finger's current position. The sound's position (in 3D) was relative to the finger's position. So if there was an object on the top right of the finger, the associated audio would sound as if it comes from the top right of the user.

The 3D positional audio navigation mode was initiated by triple tapping one of the fingers and stopped when either the user lifted the finger or they triple tapped their other finger initiating 3D positional audio relative to that finger. We wondered if receiving audio and haptic feedback for the same finger could be confusing so we allowed the user to turn the 3D positional audio off temporarily by triple tapping the active finger when receiving haptic feedback—it resumed when the haptic feedback stopped.

In the second technique, stereo audio was generated for all objects that intersected the *scanline* between the two fingers touching the screen. Thus if there was an object between the two touch points then the user would hear its associated sound. This audio was positioned relative to the mid point of the scanline. The use of the scanline was suggested by how blind users read Braille or use a quick horizontal scanning to discover the objects in a tactile graphic [17, 24] The scanline navigation mode was initiated by tapping both fingers and stopped by lifting one of the fingers from the screen. Triple tapping could also be used to temporarily turn it off.

We were not sure how effective these two navigation modes would be and so we conducted a second usability study to investigate this. The study was similar to our first study. We used graphics with 2-4 geometric shapes like the graphic in Figure 3. One shape in each graphic was significantly larger than the other shapes. Different colours were used for object boundaries, interiors and vertices. This time we associated the name of an object's geometric shape, i.e. circle, triangle, line or rectangle, with the object and this was read out when the object was queried.

For each of the two navigation modes (3D positional audio and scanline) the following two-step evaluation procedure was carried out. First we presented the participants with training graphics one at a time for that mode, which was initially on. In this part we told them which shapes were on the screen and helped them to use the mode to navigate through the shapes. We also taught them how to turn the navigation mode on and off. Second, the participant was shown one experimental graphic at a time and asked to explore the graphic and to let us know when they were ready to answer the questions. They were then asked to answer three questions about the shapes in the graphic:

1. How many objects are there in the graphic?
2. What kind of geometric shape is each object?
3. Which is the largest shape?

The time taken to initially explore the graphic and then answer each question was recorded as were their answers.

We used 6 participants in the study, some of whom had completed the first experiment. For those who had not done the first study, we had an additional training session for the haptic interaction.

Audio feedback combined with different sounds for each shape allowed participants to quickly obtain an overview of the graphic and after a first scan in most cases they correctly inferred the number of graphic elements. We found there was a slight performance benefit for the 3D positional audio mode and that there were very few errors for either mode. While participants successfully used the twiddling gesture to query objects, two of them complained that twiddling was difficult to use. All participants kept audio feedback turned on for both navigation modes, with only one person turning it off temporarily.

As expected, the scanline method was used to get an overview of the graphic. Interestingly, some of the participants also used it to get the size of the shape rather than using the haptic feedback. They started the scanning at the top with the widest scanline and narrowed the scanline to the left or to the right depending on which object they wanted to see. When they felt a haptic feedback from the vibrator motors they knew that they had touched the edges of the shape and so they could estimate the width of the shape. After this they went up and down with both fingers to find out the height of the shape. This was quite effective.

The preferences were split evenly between the two navigation modes and 4 of the 6 participants suggested that we provide both. Support for providing both also came from observation and comments by the participants suggesting that the modes were complementary: the scanline being most suited to obtaining an initial overview of the graphic and the 3D positional audio being suited to finding particular graphic elements.

4.4 Final Design

Based on the user evaluations and participant feedback we decided on the following design for the user interface for GraVVITAS. We allowed the user to feel graphic elements on the display with their left and right index fingers using haptic feedback to indicate when their finger was touching an element. Both 3D positional audio and scanline navigation modes were provided. These were controlled using triple taps and which mode was entered was dependent on how many fingers were touching the display when the mode was turned on. Graphic elements could be queried by either a twiddle or double tap gesture.

5 Evaluation

After finalizing the design we conducted a user evaluation designed to test whether GraVVITAS met our original design goal and could be used by our blind participants to effectively read and understand a variety of graphics. We tested this using three common kinds of 2D content that were quite different to each other: a table, a floor plan, and a line graph.

5.1 Design of the Graphics

An important factor in how easily an accessible graphic can be read is the layout and design of the graphic. In order to conduct the user evaluation we first needed to decide how to present the graphics to be used in the study. Our starting point were

guidelines developed for tactile graphics. These included guidelines developed by tactile transcribers which were quite low-level, giving advice on which textures are easily distinguishable, how thick lines need to be etc [29, 8]. We also referred to the higher-level design principles developed for touch screen applications with a static tactile overlay by Challis and Edwards [5]. Based on these we proposed some general principles for designing graphics for use with GraVVITAS.

The first principle was that the layout of the accessible graphic should preserve the basic structure and geometry of the original visual graphic. This was to ensure functional equivalence between the two representations and corresponds to the foundation design principle of Challis and Edwards that “A consistency of mapping should be maintained such that descriptions of actions remain valid in both the visual and non-visual representations.”

However, this does not mean that the design of the accessible graphic should exactly mirror that of the original graphic. One reason for this is that the resolution of touch is much less than sight, and so tactile graphics need to be cleaner and simpler than the original graphic. This is even more true for graphics viewed with GraVVITAS because it is difficult to distinguish objects smaller than about 5mm. Thus our second design principle was that the shapes should be simple and readily distinguishable at a 5mm resolution.

In tactile graphics the height of the tactile object is often used to distinguish between different kinds of elements, similarly to the use of colour or style in visual graphics. In the case of GraVVITAS, the choice of vibration level is the natural analogue. We determined that users could distinguish three different levels. Our design principle was that: *the vibration level should be used to distinguish different kinds of elements, with the same level used for similar kinds of objects.*

Blind users often find it difficult when encountering an unfamiliar kind of tactile graphic to gain an understanding of its structure and purpose. One of Challis and Edwards’ principles was that the design should “whenever possible encourage a specific strategy for the exploration of a particular (kind of) display.” Reflecting this principle we developed the following generic strategy for reading a graphic with GraVVITAS.

We provided at the top left corner of each graphic a “summary” rectangular shape which, when queried, would provide a short spoken description of the graphic’s purpose and content (without giving specific answers to the questions used in the usability study). For consistency we decided that the summary shape should have the same audio sound associated with it in all graphics, making it easier for the user to identify and find it.

Our suggested reading strategy was to first use scanline navigation to traverse the graphic from the top of the screen to the bottom to obtain an overview of the elements. Then to use the 3D positional audio navigation to find the summary rectangle, and use the query gesture to hear the summary. Then repeatedly to use 3D positional audio to navigate through the graphic to find the other elements. And, for each element, using the query gesture to find what each element is and to use haptic feedback to precisely locate the element and understand its geometric shape.

The other aspect we had to consider in the presentation was the design of the audio feedback provided in the navigation mode. The human perceptual subsystem groups audio streams by using different characteristics of audio such as frequency, amplitude,

temporal position, and multidimensional attributes like timbre, and tone quality [7]. Humans can differentiate about five or six different simultaneous sounds. Thus, we felt that associating audio with all elements in a complex graphic would end up being quite confusing. Instead we decided to associate audio feedback with those graphic elements that were particularly important (possibly emphasized in the original visual graphic) and objects that were natural navigational landmarks. Of course if an object had no associated audio it still has haptic feedback associated with it. We chose to use the same audio for the same kind of objects.

Using these guidelines we designed the three example graphics shown in Figures 4, 5, and 6 for the usability study. Note that the red square at the top left corner of each graphic is the summary rectangle.

For the table, the cells were represented as squares and aligned in rows and columns. We did not associate audio with the cells because we thought the regular layout of a table would make navigation straightforward. Querying a cell gave its value as well as the name of the row and column it was in. We used different vibration levels to differentiate between row headers, column headers and cells. We used thin lines to connect the headers and the cells so that it would be easier to find the neighbouring cells. The table gave the average distances ran by three different runners in three different months. We asked the following questions:

- (T1) Who ran the maximum distance in February?
- (T2) What is the distance ran by John in March?
- (T3) How was the performance of Richard?

For the floor plan we used audio feedback for the doors but not for the rooms. The idea being that this would aid understanding how to “walk” through the floorplan. The rooms were represented with filled rectangles which had two different vibration levels corresponding to their border and interior, and the doors had one strong vibration level. The doors and the rooms also had associated text information that could be queried. The floor plan was of a building with one entrance and seven rooms connected by six doors. We asked the following questions:

- (F1) Where is room 4?
- (F2) How do you go to room 7 from the entrance?
- (F3) How many doors does room 6 have?

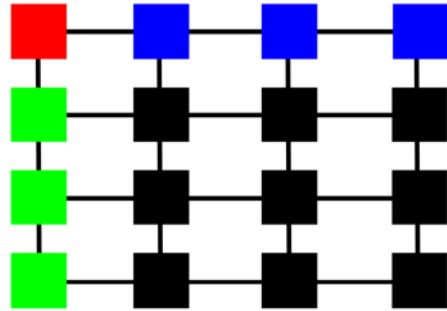


Fig. 4. Table

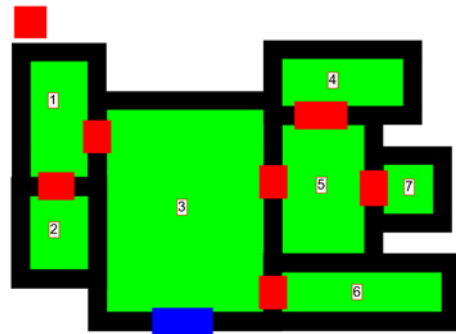


Fig. 5. Floor Plan – room numbers are not shown in the actual graphic

For the line graph the axes and labels were represented as rectangles which have their value as the non-geometric information. The lines in the graph belong to two datasets so they had different vibration levels. Small squares were used to represent the exact value of a line at a grid point. Their non-geometric information was the name of the dataset and their value on the horizontal and vertical axis. These squares also had audio associated with them so that the user could hear them while using the 3D positional mode. The line graph showed the average points scored by two different basketball teams during a seven month season. We asked the following questions:

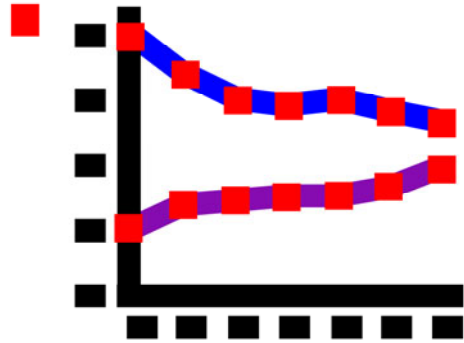


Fig. 6. Line graph

We asked the following questions:

- (L1) What is the average score of Boston Celtics in September?
- (L2) Have the Houston Rockets improved their score?
- (L3) Which team generally scored more during the season?

5.2 Usability Study

We used 6 participants all of whom had completed the second usability study. All had spent at least 4 hours using variants of the tool before the study. The primary purpose of the study was to determine if they could successfully use GraVVITAS to answer the questions about the three kinds of graphic. A secondary purpose was to obtain feedback about the drawing conventions and the interface of GraVVITAS.

We did the following for each kind of graphic: table, floor plan, and line graph. First we presented the participant with an example graphic of that kind on GraVVITAS, walking them through the graphic so as to ensure that they understood the layout convention for that kind of graphic and were comfortable using the tool. Then we presented the experimental graphic and asked them to explore it and answer the three questions about it. We recorded the answers as well as the time to answer the questions. After presenting the three kinds of graphics we asked for feedback about the tool.

All 6 participants were able to read the example graphics and answer most of the questions correctly – two incorrect answers for F2. Participant P3 could not understand the table because of the lines connecting the cells. As a result of feedback from P3 we removed the lines from the table graphic for the remaining three participants to avoid possible confusion. Question F2 was answered incorrectly by two participants because they became confused by the geometry of the floorplan.

In Table 1, we give the time in seconds taken by each participant to answer each question and the median time. The initial exploration took only a few seconds. The times vary considerably between participants. In part this is because we had not told participants to hurry and so they often checked and rechecked their answer or simply spent time “playing” with the graphic in order to better understand how to use GraVVITAS. With more experience one would expect the times to significantly reduce.

Table 1. Time taken in seconds to answer each question for the three kinds of graphic

Participant	Table			Floorplan			Line graph		
	T1	T2	T3	F1	F2	F3	L1	L2	L3
P1	67	40	45	110	1058	49	266	301	393
P2	462	45	37	50	420	300	142	80	70
P3	n/a	n/a	n/a	120	300	285	326	242	219
P4	100	92	36	62	210	360	350	158	141
P5	113	20	78	102	370	225	80	131	29
P6	121	16	35	55	388	155	180	96	159
Median	113	40	37	82	379	255	223	145	150

All participants said they liked the tool and said that with enough training they would be more than comfortable using the tool. The error and timing data, backed by participant comments, suggests that 5 out of 6 participants found the floorplan the most difficult graphic to understand, followed by the line graph, and then the table. This is not too surprising: one would expect that graphics with a more predictable layout structure are going to be easier to read by blind people.

Most participants used a reading strategy similar to the one we suggested. 4 of them started with moving a scanline from the top of the graphics to the bottom so that they could determine the location of the components. They then used one finger with 3D audio navigation mode to find the exact location of each component. When they found a component (indicated by vibration) they almost always used the query gesture to get its associated information. They repeated this process for each component.

Usually the first component they looked for was the summary object which they queried once. 4 of the participants queried this summary component a second time during the interaction but none of them a third time. 2 of the participants started by placing their fingers in the middle of the graphic and querying the objects, but later decided to query the summary shape so as to perform a more systematic exploration.

5 of the participants used the 3D audio all the time, only 1 of them turned it off saying that s/he could remember where each component was. When reading the line graph, 5 of them used two fingers to answer the trend question, and 1 of them preferred to read each individual data point.

3 of the participants had problems with double tapping to query an object because our implementation required both taps to intersect with the object and if the user was tapping on the border of the object then they were quite likely to miss the object on the next tap, meaning that the tool would not provide the expected query information. Several participants suggested that rather than having to explicitly query an object the associated audio description should be triggered when the user first touches the object. This seems like a good improvement. Other suggestions were to provide more meaningful audio with objects for the navigation mode.

6 Conclusion

We have described the design and evaluation of a novel computer tool, GraVVITAS, for presenting graphics to people who are blind. It demonstrates that touch-screen

technology and haptic feedback devices have now reached a point where they have become a viable approach to presenting accessible graphics. We believe that in the next few years such an approach will become the standard technique for presenting accessible graphics and other two dimensional information to blind people, much as screen readers are now the standard technique for presentation of textual content. While touch screens and SVG are still not widely used, we believe that in a few years they will be mainstream technology.

We had three design requirements when designing GraVVITAS. The first was that it could be used effectively by people who are blind to read an accessible version of a wide range of graphics and 2D content. Our user studies provide some evidence that this is true, allowing the participants to answer questions about different kinds of graphics. We observed that in all of the user studies the participants referred to the shapes in terms of their relative position to each other and their position overall in the graphic. This provides additional evidence that the tool allows the blind user to build an internal representation that is functionally similar to that of the sighted user looking at the original graphic. A limitation of the current evaluation is its small size and that the same participants were used in several studies. In the future we plan to conduct an evaluation with a larger set of participants.

The second design requirement was that the tool is practical. The tool is inexpensive to buy and to operate: it was built from the off-the-shelf components with a total cost of US \$2,508 of which nearly \$2,000 was for the Dell Latitude XT Tablet PC (with multi-touch screen). It costs virtually nothing to operate for the end user although there are still time costs involved in the creation of the graphics for the effort of transcribers. Its size and design mean that it could be used in classrooms, home and work environments, and it could be integrated with other applications such as DAISY to read graphics contained in books etc. The main limiting factor is that it currently requires a human to produce the accessible graphic. This can be done using freely available SVG editors such as Inkscape so is an improvement on the need for access to special purpose printers and paper when producing tactile graphics. However, our longer term goal is to automate this process, along the lines of TGA, by automatically generating accessible graphics from SVG.

The final design requirement was that the tool supports interactive, active use of graphics. In principle because the display is refreshable GraVVITAS supports this. However the current software does not yet take advantage of this. We plan to explore how best to provide user-guided zooming and panning. This will help us overcome the low resolution of the graphics displayed on the screen. We will also explore how to support creation of graphics by blind users through applications like graphing calculators and also with a generic graphics authoring tool.

Finally, we would like to further investigate the design of our system. First we want to examine how non-blind users in limited situations can also benefit from it [3]. Second, we want to explore whether blind users with concerns about wearable systems can prefer to use other devices such as electro-vibration surfaces where the feedback is produced on the device.

Acknowledgements. We thank the participants for their effort and great feedback. We also thank the Vision Australia transcription department, especially Peter Szikla, for their support for the project and user studies. We thank Michael Wybrow for his insightful comments. We also acknowledge the support of the ARC through

Discovery Project Grant DP0987168. Finally, we thank Chongqing Linglong Electronics Co.,Ltd. (<http://en.lnlon.com>) for providing the vibrator motors.

References

- [1] Bau, O., Poupyrev, I., Israr, A., Harrison, C.: TeslaTouch: electrovibration for touch surfaces. In: Proc. of the UIST, pp. 283–292. ACM, New York (2010)
- [2] Benali-Khoudja, M., Hafez, M., Alexandre, J., Kheddar, A.: Tactile interfaces: a state-of-the-art survey. In: Int. Symposium on Robotics (2004)
- [3] Buxton, B., Buxton, W.: Sketching user experiences: getting the design right and the right design. Morgan Kaufmann, San Francisco (2007)
- [4] Cattaneo, Z., et al.: Imagery and spatial processes in blindness and visual impairment. *Neuroscience and Biobehavioral Reviews* 32(8), 1346–1360 (2008)
- [5] Challis, B., Edwards, A.: Design principles for tactile interaction. In: Brewster, S., Murray-Smith, R. (eds.) *Haptic HCI 2000*. LNCS, vol. 2058, pp. 17–24. Springer, Heidelberg (2001)
- [6] Coren, S., Ward, L., Enns, J.: *Sensation and perception*. Wiley, Chichester (2004)
- [7] Deutsch, D.: *The psychology of music*. Academic Pr., London (1999)
- [8] Edman, P.: *Tactile graphics*. American Foundation for the Blind (1992)
- [9] Elzer, S., et al.: A probabilistic framework for recognizing intention in information graphics. In: Proc. of the Int. Joint Conf. on AI, vol. 19, pp. 1042–1047 (2005)
- [10] Eriksson, Y.: *Tactile Pictures: Pictorial Representations for the Blind 1784-1940*. Gothenburg Uni. Press (1998)
- [11] Foulke, E.: Reading braille. In: *Tactual Perception*, pp. 223–233. Cambridge Uni. Press, Cambridge (1982)
- [12] Frauenberger, C., Noisternig, M.: 3D Audio Interfaces for the Blind. In: *Workshop on Nomadic Data Services and Mobility*, pp. 11–12 (2003)
- [13] Gardner, J., Bulatov, V.: Scientific Diagrams Made Easy with IVEO. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) *ICCHP 2006*. LNCS, vol. 4061, pp. 1243–1250. Springer, Heidelberg (2006)
- [14] Goncu, C., Marriott, K., Aldrich, F.: Tactile Diagrams: Worth Ten Thousand Words? In: Goel, A.K., Jamnik, M., Narayanan, N.H. (eds.) *Diagrams 2010*. LNCS, vol. 6170, pp. 257–263. Springer, Heidelberg (2010)
- [15] Goncu, C., Marriott, K., Hurst, J.: Usability of Accessible Bar Charts. In: Goel, A.K., Jamnik, M., Narayanan, N.H. (eds.) *Diagrams 2010*. LNCS, vol. 6170, pp. 167–181. Springer, Heidelberg (2010)
- [16] Hatwell, Y.: *Images and Non-visual Spatial Representations in the Blind*, pp. 13–35. John Libbey Eurotext (1993), <http://books.google.com>
- [17] Hatwell, Y., Martinez-Sarrochi, F.: The tactile reading of maps and drawings, and the access of blind people to works of art. In: *Touching for Knowing: Cognitive Psychology of Haptic Manual Perception*. John Benjamins B.V., Amsterdam (2003)
- [18] Kensing, F., Blomberg, J.: Participatory design: Issues and concerns. *Computer Supported Cooperative Work (CSCW)* 7(3), 167–185 (1998)
- [19] Ladner, R., et al.: Automating tactile graphics translation. In: Proc. of the 7th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 150–157 (2005)
- [20] Landau, S., Gourgey, K.: Development of a Talking Tactile Tablet. *Information Technology and Disabilities* 7(2) (2001)

- [21] Larkin, J., Simon, H.: Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science* 11(1), 65–100 (1987)
- [22] Manshad, M., Manshad, A.: Multimodal Vision Glove for Touchscreens. In: *Proc. of ACM ASSETS*, pp. 251–252. ACM, New York (2008)
- [23] McGookin, D., Brewster, S.: MultiVis: Improving Access to Visualisations for Visually Impaired People. In: *CHI 2006 Extended Abstracts*, pp. 267–270. ACM, New York (2006)
- [24] Millar, S.: *Reading by touch*. Routledge, New York (1997)
- [25] Millar, S.: *Space and sense*. Psychology Press/Taylor & Francis (2008)
- [26] Nagasaka, D., et al.: A Real-Time Network Board Game System Using Tactile and Auditory Senses for the Visually Impaired. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) *ICCHP 2010*. LNCS, vol. 6179, pp. 255–262. Springer, Heidelberg (2010)
- [27] Petrie, H., et al.: TeDUB: A System for Presenting and Exploring Technical Drawings for Blind People. In: Miesenberger, K., Klaus, J., Zagler, W.L. (eds.) *ICCHP 2002*. LNCS, vol. 2398, pp. 537–539. Springer, Heidelberg (2002)
- [28] Petrie, H., Hamilton, F., King, N., Pavan, P.: Remote usability evaluations with disabled people. In: *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1133–1141. ACM, New York (2006)
- [29] *Tactile Diagram Manual*, Purdue University (2002)
- [30] Scaife, M., Rogers, Y.: External cognition: how do graphical representations work? *International Journal of Human-Computer Studies* 45(2), 185–213 (1996)
- [31] Shimojima, A.: The graphic-linguistic distinction: Exploring alternatives. *Artificial Intelligence Review* 13(4), 313–335 (1999)
- [32] Takagi, H., Saito, S., Fukuda, K., Asakawa, C.: Analysis of navigability of Web applications for improving blind usability. *ACM Trans. on Computer-Human Interaction* 14(3), 13 (2007)
- [33] Tversky, B.: Spatial schemas in depictions. In: *Spatial Schemas and Abstract Thought*, pp. 79–111. MIT Press, Cambridge (2001)
- [34] Ungar, S.: Cognitive Mapping without Visual Experience. In: *Cognitive mapping: Past, Present and Future*, pp. 221–248 (2000)
- [35] Velleman, E., van Tol, R., Huijberts, S., Verwey, H.: 3D Shooting Games, Multimodal Games, Sound Games and More Working Examples of the Future of Games for the Blind. In: Miesenberger, K., Klaus, J., Zagler, W.L., Burger, D. (eds.) *ICCHP 2004*. LNCS, vol. 3118, pp. 257–263. Springer, Heidelberg (2004)
- [36] Vidal-Verdu, F., Hafez, M.: Graphical Tactile Displays for Visually-Impaired People. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 15(1), 119–130 (2007)
- [37] Wall, S., Brewster, S.: Feeling What You Hear: Tactile Feedback for Navigation of Audio Graphs. In: *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1123–1132. ACM, New York (2006)

Designing a Playful Communication Support Tool for Persons with Aphasia

Abdullah Al Mahmud, Idowu I.B.I. Ayoola, and Jean-Bernard Martens

Department of Industrial Design, Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
{a.al-mahmud, j.b.o.s.martens}@tue.nl

Abstract. Many studies have investigated ways to leverage communication with people with aphasia. Here, a new concept is developed for people with non-severe aphasia in a way that accesses the emotional and unaware layer of a conversation and then communicates certain information to the partner hence; introducing new dynamics and structure to a conversation. We present the concept with detailed design and expert evaluation results.

Keywords: Aphasia, Accessibility, Storytelling, Contextual interview, Assistive technology.

1 Introduction

Aphasia is an acquired communication disorder that impairs a person's ability to process language or to understand others often caused by brain injury or trauma. Aphasia affects language comprehension and generation, such that people's ability to express themselves verbally suffers [3]. People with Broca's aphasia or expressive aphasia usually can understand or read what other people say or write but they have problems in expressing themselves verbally and in writing. The consequence of aphasia is that people have problems maintaining contacts with their friends cannot participate in social exchange and eventually become passive and socially isolated.

There has been a growing interest in improving the quality of life for people with aphasia through technology intervention [1, 2]. It has been observed that the use of assistive technology can reduce social isolation and improve independence. Therefore, new tools are emerging for specific purposes such as helping aphasics while cooking [6], using internet [1] etc. These applications support higher-level communication needs of aphasics. However, people have several other communication needs such as social closeness and information transfer [5]. Up to now augmentative and alternative communication (AAC) devices have focused mainly on helping aphasics with basic communication needs [4]. Though these AACs are mostly used during therapy period they are however not conversation tools. Therefore, we see an opportunity in assisting aphasics to participate in conversations by our proposed design.

In this paper we focus on the development of a supportive device to help leverage communication with a target group of aphasic people besides using online strategies. We explore possibilities for a new communication support tool and present a final

design that creates a new layer of interaction within a conversation and hopes to reconstruct the flow of information and enhance the overall experience derived from a dialogue in an unobtrusive way. This was achieved with the use of a monitored stress ball to stimulate an unaware interaction and then providing a constructive feedback to the partner. An iteration approach is observed to investigate the topic and to further develop into a final design. In understanding the problem domain, the researcher began with literature research to know ways of aiding communication or social activities for people with Aphasia. Furthermore the psychological needs for the target groups were briefly identified to settle on an approach for creating a new product for them. Lastly, an experiment was conducted to observe the kind of behaviors that may emerge in a group with limited communication abilities similar to aphasia. Certain design interventions were derived from these investigations and a final design opportunity was selected for development.

2 Understanding the Problem Domain

At the start of the project we investigated several questions such as a) How can we aid communication or social activities for aphasic patients? b) Can a group of people with impaired communication problems perform a common goal and how? An experiment designed for 15 participants was conducted to answer the questions. The objective was to make people work together using a square game by allowing them to perform physical tasks that oblige them to communicate and cooperate. All participants were students in their early twenties and had no impaired ability however they were not allowed to talk during the experiment depending on the type they belonged.

Each participant was allocated as member of type “A” or “B”. Members of type “A” were not allowed to talk throughout the activities while those in type “B” were allowed to talk. The experiment was divided into three groups with sorted members. Group 1 had five members belonging to type “B”. Group 2 had three members of type “A” and two of type “B”. Group three had four members as type “A”. The table below summarizes the division.

Table 1. Group division and number of members

<i>Group ID</i>	<i>Members of type “A”</i>	<i>Members of type “B”</i>
Group 1	-	5
Group 2	3	2
Group 3	4 (omitted one person)	-

After experimenting, a questionnaire was handed out to members of groups to obtain both quantitative and qualitative data. The questions emphasized on their experience in the group as different types and to know how they improvised in order to communicate with other members. Also, by observing a video of the experiment, more information was obtained. Results from the experiment shows that even if members of a group is not allowed to talk, the group can still perform certain activity together given a common goal. Members are able to work together using

gestures, pointing, etc, to communicate with one another and still retain a level of involvement/satisfaction. Eye contact was hardly sustained especially with people that were not allowed to talk. As a result, communication between members was shallow but rather focused their attention on the given task. Furthermore, participants who were not allowed to talk appeared less dominant hence, the social cohesion within the group may be improved by providing clues to help them properly take turns in a dialogue. Pointing gestures was the most effective method used to communicate especially with members who were not allowed to talk. In chapter 3, these insights are developed into actual concepts.

3 Design Interventions and Development

An idea generation session was conducted to obtain more ideas from certain directions. Below are three design opportunities that were identified and the third was selected to be further developed.

Recap. This idea attempts to bring an aphasic in contact with others to talk about their daily experiences. A sound recording device is used to record sound in a daily context and later replayed to stimulate discussion with friends or family. At the instance of recording the sound, pictures related to the subject matter or context can be captured and are automatically tagged to the sound sample to support future discussion when sharing. Figure 1(top) illustrates this idea; two individuals used the device to record their daily activity and later met with others for social discussions.

Information Schema. In this concept a person with aphasia may re-create or organize photos / illustrations on a timeline to generate or represent stories which may support future communication with others as illustrated on figure 1(middle). By pointing at the pictures on the board, he may be able to tell his stories easily without having to recall the words. The picture shows an aphasic creating his story on his playground, and later used them to tell his story to a friend.

The mediator concept. The design opportunity here is to enhance the flow or contributions of members in a conversation that may be between a person with aphasia and his partner. The partner in this case can be a therapist, family or friend. Having identified “turn taking” as an important aspect of communication, the concept attempts to provide feedback to partners in order to establish a convenient moment to interrupt or give clues. This is conceptualized with the use of objects that affords tinkering like in a stress ball in order to cause subconscious interaction with the object. According to the mental activity or state of a person, their interaction with the object changes and this information can be abstracted and used to provide feedback to the partner. Hopefully, this information flow would introduce a new dynamics/structure into their conversation. Figure 1(bottom) illustrates two people talking and holding different objects each. Through the objects, partners can receive impulse from the other and may as well initiate certain impulse to the other.

4 Final Concept

The design opportunity further developed is the mediator concept because it provides a rare opportunity in design for aphasics. It deviates from the traditional photo and sound often capitalized or abused within this context however, it is aimed to focus on a rich and interactive experience that may be derived from a dialogue. The following subtopics present the steps of developments.

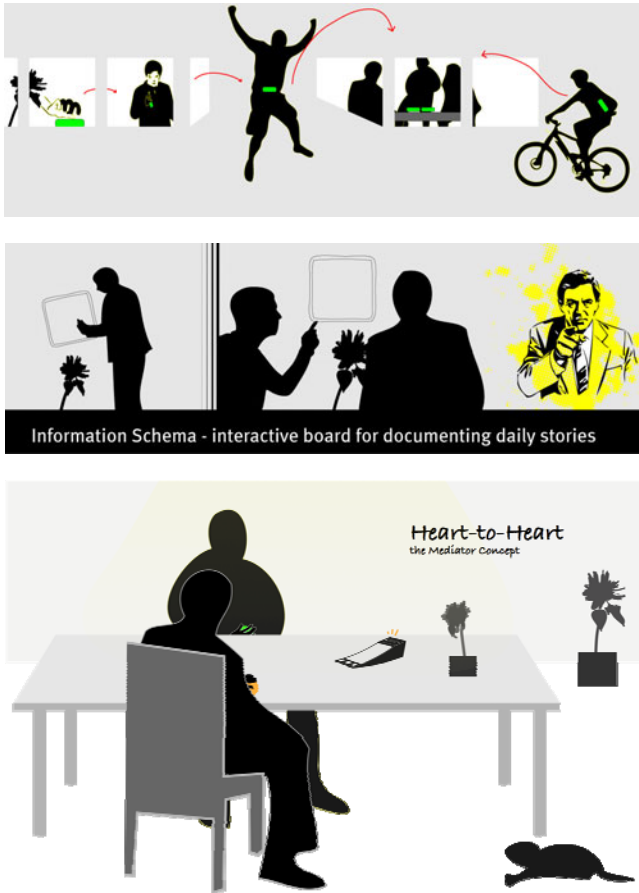


Fig. 1. Scenario of Audio Recap concept (top), scenario of Information Schema Recap concept (middle), scenario of Information Schema Recap concept (bottom)

A user test and concept evaluation was made to know how people unconsciously interact with objects in their hands and how much this interaction changes depending on the state on thinking. The prototype used for this investigation consists of two balloons inter-connected using a plastic tube. When one end is pressed, the pressure at the other end increases causing an instant feedback to the person holding at the other end. The first test was with two people, where one asked questions to the other.

Different questions were asked and the difficulty level of the questions varied. It was observed that people unconsciously play with the balloon when they are relaxed. However, when the question demands more concentration, they momentarily stop playing with the ball.

4.1 Material Explorations

Throughout the design process, various materials were explored for building the ball. This is important to select the best material that would trigger the kind of interaction we need and to provide optimum comfort for the user. Examples of the materials created are silicon ball, balloon filled with powder, balloon filled with tinny wooden balls, tightly dense stress ball, PU ball, etc. PU ball was preferred over other materials because it allows full compression yet does not strain the hands. It is a perforated material and allows air to escape when pressed and then return to its original shape immediately after release. However due to unavailability of this material, silicon material was used. The silicon material provided similar affordances however do not allow full compression, as it is airtight.

4.2 Digitizing Interaction and Feedback

A silicon ball was made and monitored through the change in air pressure when pressed. A vibrator motor was used to provide instant feedback according to the results of the analysis. The prototype provides feedback in a way that makes the device seem invisible to users. Different feedback modalities were investigated however, haptic feedback was quickly settled upon in order to avoid visual distractions and to retain eye contact during the conversation. From empirical analysis, feedback in the rhythm of heartbeat was preferred over continuous vibration as it may create empathy amongst the partners. In essence, the harder the ball is pressed, the stronger the heartbeat, and when the user freezes while squashing the ball, the ball beats faster. A pilot test was done with some design students to ensure the prototype was robust enough for future testing and evaluations.

4.3 Introducing a Desktop Interface

A desktop interface (Fig. 2 top-left) is introduced and prototyped as part of the concept in order to provide visual cues to support the conversation. An example of the programs that may be installed may be a photo album organized in events. This way when a conversation began, a photo album can be opened to provide picture support. Another example may be a Phonetic & Alphabet Chat to be used to scramble letters or sounds that may serve as clues. The purpose for linking the handheld device and the desktop device is to strengthen the use of computers as visual or auditory cues. The handheld device provides information abstracted from the conversation in order to invite users to use the desktop device when needed. This may cause the desktop device and similar other computer applications built for support to have better utilitarian purpose.

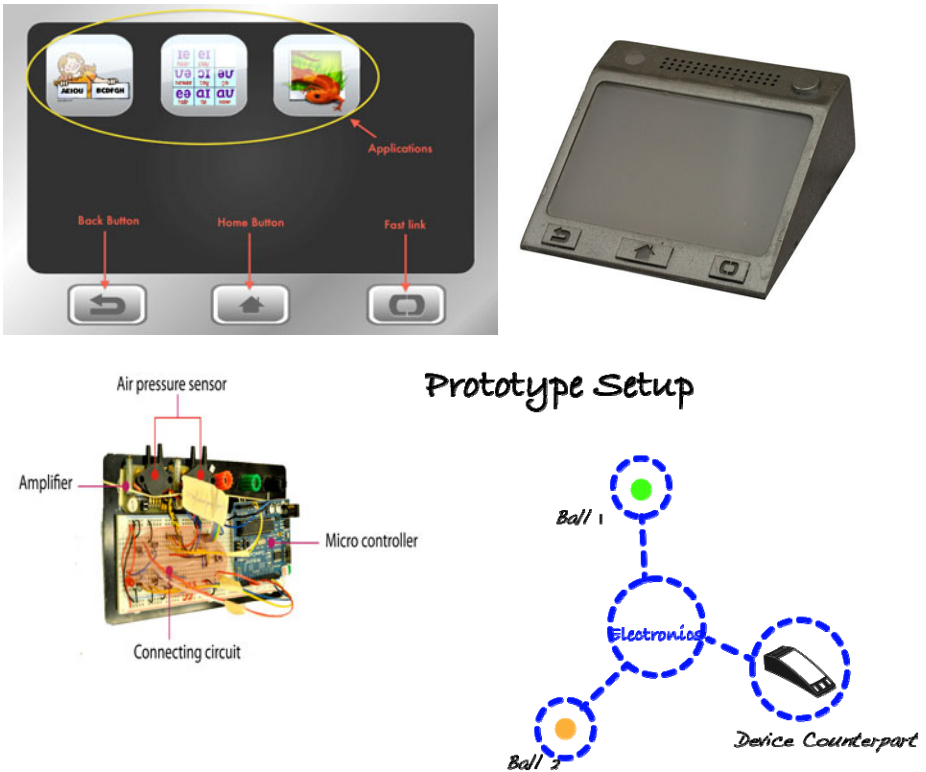


Fig. 2. The Desktop interface (top-left), the form of the desktop tool (top-right), the electronic prototype (bottom-left, and the prototype setup with balls (bottom-right)

5 Expert Evaluation

Having created a preliminary prototype to evaluate the design unity and rationale, a discussion session with an interaction expert was organized. The preliminary prototype was used to experience the design. The evaluation took place in an open office area. The working prototype was placed on a desk together with the desktop device. A computer was used to show the interface of the desktop device. Two chairs were placed closely apart to motivate a close proximity conversation. The goal was to enable the expert to experience the prototype in order to evaluate the rationale of the concept. This is aimed to inspire new insight for the design development.

Results: Interaction with objects of the design is classified into aware and unaware layers. The expert identifies interaction with the ball devices as unaware layer of communication as it is non-obtrusive to the users but yet would influence them. He believes it would change the quality of their conversation. However, when attention moves to the desktop device, the interaction with the system changes phase and causes the devices to become visible. He recommends developing the interaction and

feedback with the ball devices because it may change people's expressive attitude and satisfaction in a dialogue.

- i. The feedback received in form of heartbeat may have positive influence on users because it initiates the feeling of contact without physical touch. This may cause the partners to become more connected both physically and emotionally. He thinks there might be information overload if the person with aphasia regularly receives dynamic response from the device. This reveals the potential need to provide the users with some control; i.e., to choose to receive dynamic feedback from the device. His advice is to investigate the potentials of this design firstly within a therapeutic or structured conversation before moving on to general applications.
- ii. The expert points out that users are constrained in movement especially because of the visible wires/connections. This may have psychological effect on users and the designed effect may be limited.

6 Redesigning the Final Prototype: Mediator Concept

The final prototype incorporates a wireless module and two ball devices for initiating and receiving impulse. It does not include the desktop device because the current line of development focus on the interaction and influence with the ball devices. Figure 3 is an overview of the final design. The wireless module includes a ball device and a case to hold the extended part of the electronics. It is implemented in a decentralized way, which makes it more sustainable, or standalone. On the casing is a button and a light source which shows the current state of the device. There are two selectable modes implemented on each module, one allows full functionality and the other allows the user to deactivate dynamic impulse meaning they would not receive regular impulse from the partner.



Fig. 3. The final prototype

Each module includes an air pressure sensor that measures the relative pressure in the ball to the atmospheric pressure. This signal is amplified within a range of 0-3.3 volts for analogue to digital conversion on an ATmega328 micro-controller chip. An interval routine that runs on the micro-controller processes this signal and outputs a

modulation scheme/command through a wireless protocol to the second module. An interrupt routine is called when data arrives at the wireless port. Depending on the current mode of the device, the data is used to modulate the vibration motor embedded in the ball to stimulate haptic feedback.

7 Conclusion and Future Steps

Certain design interventions were derived and a course of development was settled upon which introduces a new dimension of a product for people with aphasia. The prospective product is a pair of wireless modules, which includes a ball in form of a stress ball, and a casing that holds parts of the electronics and controls. The end product was designed to access the unaware layer of a conversation in order to provide partners with a constructive feedback that may enrich the social experience of users.

Prototypes were created to validate certain design steps and later developed into a more experiential and advanced prototype, momentarily as a final design. Further investigations should be made to quantify the influence of the new product within a conversation. This would provide more insight for developments. In future, PU rubber material should be adopted, as it would enhance the interaction with the ball. Furthermore, investigations should be made to know if people are more likely to use a desktop application when the ball device is introduced. At this point, it is essential to evaluate the prototype with an aphasia therapist/expert to gain contextual insight; afterwards, a user test with actual users like aphasics may commence. An initial test between a person with aphasia and his/her therapist should be done before trying out in more dynamic situations. The design concept developed in this paper can eventually serve as an intervention tool, which can be integrated into application interfaces already built for people with aphasia.

References

- [1] Egan, J., Worrall, L., et al.: Accessible internet training package helps people with aphasia cross the digital divide. *Aphasiology* 18(3), 265–280 (2004)
- [2] Koppenol, T., Al Mahmud, A., Martens, J.-B.: When words fall short: helping people with aphasia to express. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) *ICCHP 2010*. LNCS, vol. 6180, pp. 45–48. Springer, Heidelberg (2010)
- [3] Hillis, A.: Aphasia: progress in the last quarter of a century. *Neurology* 69(2), 200–213 (2007)
- [4] Hux, K., Manasse, N., Weiss, A., Beukelman, D.: Augmentative and alternative communication. In: Chapey, R. (ed.) *Language Intervention Strategies in Adult Aphasia*, pp. 675–689. Williams & Wilkins, Baltimore (2001)
- [5] Light, J.: Interaction involving individuals using augmentative and alternative communication systems: State of the art and future directions. *Augmentative and Alternative Communication* 4(2), 66–82 (1988)
- [6] Tee, K., Moffatt, K., et al.: A visual recipe book for persons with language impairments. In: *Proc. CHI 2005*, pp. 501–510. ACM, New York (2005)

How to Make Numerical Information Accessible: Experimental Identification of Simplification Strategies

Susana Bautista¹, Raquel Hervás¹, Pablo Gervás²,
Richard Power³, and Sandra Williams³

¹ Universidad Complutense de Madrid, Spain

² Instituto de Tecnología del Conocimiento, Madrid, Spain

³ Department of Computing, The Open University, Milton Keynes MK76AA, UK
{subautis, raquelhb}@fdi.ucm.es, pgervas@sip.ucm.es,
{r.power, s.h.williams}@open.ac.uk

Abstract. Public information services and documents should be accessible to the widest possible readership. Information in newspapers often takes the form of numerical expressions which pose comprehension problems for people with limited education. A first possible approach to solve this important social problem is making numerical information accessible by rewriting difficult numerical expressions in a simpler way. To obtain guidelines for performing this task automatically, we have carried out a survey in which experts in numeracy were asked to simplify a range of proportion expressions, with three readerships in mind: (a) people who did not understand percentages; (b) people who did not understand decimals; (c) more generally, people with poor numeracy. Responses were consistent with our intuitions about how common values are considered simpler and how the value of the original expression influences the chosen simplification.

Keywords: numerical information, simplification strategies.

1 Introduction

A United Nations report [1] recommends that public information services and documents should be accessible to the widest possible readership. Information in newspapers often takes the form of numerical expressions (e.g., economic statistics, demographic data) which pose comprehension problems for people with limited education. A UK Government Survey in 2003 estimated that 6.8 million adults had insufficient numeracy skills to perform simple everyday tasks, and that 23.8 million adults would be unable to achieve grade C in the GCSE maths examination for 16-year-old school children [2].

A first possible approach to solve this important social problem of making numerical information accessible is to rewrite difficult numerical expressions more simply. Such an approach would require a set of rewriting strategies yielding expressions that are linguistically correct, easier to understand than the original, and as close as possible to the original meaning. For example, ‘25.9%’ could be rewritten as ‘just over a

quarter'. Simplification may in some cases entail loss of precision, but this is not necessarily a bad thing, for several reasons. Loss of precision can be signaled linguistically by numerical hedges such as 'around', 'more than' and 'a little under', so it need not be misleading. As Krifka has argued, competent writers and speakers frequently approximate numerical information and readers and hearers can readily recognize this, even when no hedge is present, especially when numbers are round [3]. For instance, in 'the distance from Oxford to Cambridge is 100 miles' it is clear that 100 miles is an approximation. Williams and Power [4] showed that writers tend to approximate numerical quantities early in a document, then give more precise versions of the same quantities later. As Krifka argues in the same paper [3], an inappropriately high level of precision would flout Grice's Maxim of Quantity [5] by giving too much information. There cannot be many situations in which we need to know that the distance from Oxford to Cambridge is 100.48 miles, for example.

This paper presents an exploratory survey in which experts in numeracy were asked to simplify numerical expressions (presented in context) for several kinds of readership, with the aim of collecting a repertoire of rewriting strategies that can be applied in an automatic text simplification system.

2 Background

Text simplification, a relative new task in Natural Language Processing, has been directed mainly at syntactic constructions and lexical choices that some readers find difficult, such as long sentences, passives, coordinate and subordinate clauses, abstract words, low frequency words, and abbreviations. Chandrasekar et al. [6] introduced a two-stage process, first transforming from sentence to syntactic tree, then from syntactic tree to new sentence; Siddharthan [7] instead proposed a three-stage process comprising analysis, transformation and generation. In 1998, the project PSET [8] employed lexical as well as syntactic simplifications. Other researchers have focused on the generation of readable texts for readers with low basic skills [9], and for teaching foreign languages [10]. However, to our knowledge, there have been no previous attempts to automatically simplify *numerical* information in texts.

A corpus of numerical expressions was collected for the NUMGEN project [4]. The corpus contains 10 sets of newspaper articles and scientific papers (110 texts in total). Each set is a collection of articles on the same topic - e.g., the increased risk of breast cancer in red meat eaters, and the decline in the puffin population on the Isle of May. Within each set, identical numerical facts are presented in a variety of linguistic and mathematical forms.

3 Experiment

Candidate rewriting strategies may be obtained in two ways: one is to collect them directly from human authors, another is to validate strategies mined from a large corpus. Our experiment employs the first option.

3.1 Underlying Assumptions

In this paper we consider a ‘numerical expression’ (NE) to be a phrase that presents a quantity, optionally modified by a numerical hedge as in ‘more than a quarter’ or ‘around 97 %’. To date, we have restricted coverage to proportions - i.e., fractions, ratios and percentages. We have two working hypothesis:

- H1:** When experienced writers choose numerical expressions for readers with low numeracy, they tend to prefer round or common values to precise values. For example, halves, thirds and quarters are usually preferred to eightieths or forty-ninths, and expressions like *N in 10* or *N in 100* are chosen instead of *N in 365* or *N in 29*.
- H2:** The choice between different simplification strategies (fractions, ratios, percentages) is influenced by the value of the proportion, with values in the central range (say 0.2 to 0.8) and values at the extreme ranges (say 0.0-0.2 and 0.8-1.0) favouring different strategies.

3.2 Materials

We focused on simplification strategies at two levels: decimal percentages, and whole-number percentages. Three sets of candidate sentences were chosen from the NUMGEN corpus for presentation to participants: eight sentences containing only decimal percentages, and two sets of eight sentences containing mixed whole-number and decimal percentages. Although the number of sentences in each set was eight, the number of numerical expressions was larger as some sentences contained more than one proportion expression.

A wide spread of proportion values was present in each set, including the two end points at nearly 0.0 and almost 1.0. We also included some numerical expressions with hedges and sentences from different topics in the corpus. In short, we included as many variations in context, precision and different wordings as possible.

3.3 Participants

Our experimental evaluation involved 34 participants, considering only the ones that answered at least one question. They were primary or secondary school mathematics teachers or adult basic numeracy tutors, all native English speakers. The task of simplifying numerical expressions is difficult, but it is a task that this group seemed well qualified to tackle since they are highly numerate and accustomed to talking to people who do not understand mathematical concepts very well. We found participants through personal contacts and posts to Internet forums for mathematics teachers and numeracy tutors.

3.4 Survey Design and Implementation

Our survey took the form of a questionnaire in which participants were shown a sentence containing one or more numerical expressions which they were asked to simplify. The survey was divided into three parts as follows:

1. Simplification of numerical expressions for a person who can not understand percentages. We will refer to this part as ‘No Percentages’.
2. Simplification of numerical expressions for a person who can not understand decimals. We will refer to this part as ‘No Decimals’.
3. Free simplification of numerical expressions for a person with poor numeracy. We will refer to this part as ‘Free Simplification’.

For part (2), the set of sentences containing only decimal percentages was used. One of the two mixed sets of sentences with whole-number and decimal percentages was used for part (1) and the other for part (3). The experiment was presented on SurveyMonkey¹, a commonly-used provider of web surveys.

We asked participants to provide simplifications for numerical expressions that were marked in each sentence. Below the sentence, each numerical expression was shown beside a text box in which the participant was asked to type the simplified version. Our instructions said that numerical expressions could be simplified using any format: number words, digits, fractions, ratios, etc. and that approximators such as ‘more than’, ‘almost’ and so on could be introduced if necessary. Participants were also told that the meaning of the simplified expression should be as close to the original expression as possible and that, if necessary, they could rewrite part of the original sentence.

4 Results

The results of the survey were carefully analyzed as follows. First, within each block of questions, a set of simplification strategies was identified for each specific numerical expression. These strategies were then grouped together according to the mathematical forms and/or linguistic expressions employed (fractions, ratios, percentages). Where necessary, they were subdivided further according to choices of numerical values for the constituents of the simplified expressions (denominators in fractions, or reference value in ratios, for example). Not all simplification strategies occur with enough frequency to merit detailed analysis; the approach followed here has been to group together (under a generic label of *Others*) all simplification strategies with a low frequency of use with respect to the total (for example, in the case of fractions, a total of ten different kinds of fractions were used (hundredths, sixths, tenths, etc.), but we only represent in labeled sub-columns the ones with significant usage; the rest are summed in the *Others* sub-columns). The non-numeric column represents simplified expressions where no numbers were used like ‘almost all’ or ‘around none’. Remaining simplifications (*Rem.* column) are rewritings of the whole sentence or parts of it, coinciding with comments expressed by the participants that sometimes the whole sentence would be better understood if the non-numerical part was also simplified, and some deletions. The observed frequencies (represented in percentages) of the different simplification strategies are given in Table 1. Rows do not add up 100% as not all participants gave an answer for all numerical expressions.

¹ www.surveymonkey.com

Table 1. Frequencies of simplification strategies for 34 participants: (1) No Percentages: intended for people who do not understand *percentages*, (2) No Decimals: intended for people who do not understand *decimals*, and (3) Free Simplification: intended for people with poor numeracy. Frequencies are represented in percentages

NO PERCENTAGES (%)												
Numerical Expression	Fractions				Total	Ratios			Total	Non-numeric	Percent-ages	Rem.
	Halves	Thirds	Quarters	Others		N in 10	N in 100	Others				
more than 1%	3			15	18		6		6	15	18	24
2%				6	6		12	6	18	3	12	38
16.8%			3	24	26		15	50	65		9	
27%		9	71	3	82			12	12		6	
at least 30%		21	9	12	41	29		6	35		3	9
40%	21	6		26	53	29			29		6	3
56%	82				82						6	3
63%	24	41		9	74	9		15	24		3	
75%			32		32			29	29	3		24
97.2%				3	3	3	29	6	38	21	18	12
98%				6	6		12		12	65	3	9
Mean	12%	7%	10%	9%	39%	6%	7%	11%	24%	10%	7%	11%
NO DECIMALS (%)												
Numerical Expression	Fractions				Total	Ratios			Total	Non-numeric	Percent-ages	Rem.
	Halves	Thirds	Quarters	Others		N in 10	N in 100	Others				
0.6%	3			3	6	3	6		9	6	47	3
2.8%				3	3	24			24		47	9
6.1%						15		3	18		50	3
7.5%				12	12	3	6	3	12		50	6
15.5%				15	15	3	6	3	12		44	9
25.9%			15		15		3	9	12		38	3
29.1%				3	3	9	3	3	15		50	3
35.4%		9		3	12	9	3	3	15		41	3
50.8%	44				44		3		3		21	3
73.9%			44		44		3	3	6		18	3
87.8%				3	3	9	3	3	15		47	3
96.9%				3	3	6	3	3	12		29	12
96.9%				6	6	9	6	3	18		21	6
97.2%				3	3	6	6	6	18	3	41	6
97.2%				3	3	12	3	3	18	3	32	6
98.2%				3	3	9	3	3	15	6	44	3
Mean	3%	1%	4%	4%	11%	7%	3%	3%	14%	1%	39%	5%
FREE SIMPLIFICATION (%)												
Numerical Expression	Fractions				Total	Ratios			Total	Non-numeric	Percent-ages	Rem.
	Halves	Thirds	Quarters	Others		N in 10	N in 100	Others				
0.7%							6		6	18	9	26
12%				6	6	12	3	6	21		21	3
26%			41		41			12	12			3
36%		41			41	3		6	9			3
53%	41				41						6	6
65%	6	15			21	3	9	6	18		3	12
75%			15		15			9	9	6	3	15
91%						21	9		29	6	6	12
above 97%						3	29		32	12	6	
Mean	5%	6%	6%	1%	18%	5%	6%	4%	15%	5%	6%	9%

In order to analyse the results we performed a one-way analysis of variance (ANOVA), which results are represented in Table 2. When considering the whole survey (*Whole* column), there is no significant difference in the use of fractions, ratios and percentages. Only the use of non-numeric expressions is significant, but this is due to their low usage. However, when analysing the survey by parts we find interesting results.

Table 2. Results of ANOVA test. Strategies which do not share a letter are significantly different.

Strategy	No Percentage			No Decimals			Free Simplif.		Whole	
Fractions	A			A			A		A	
Ratios		B		A			A		A	
Percentages			C		B			B	A	
Non-Numeric			C			C		B		B

Overall, fractions are the preferred simplification for *people who do not understand percentages*. Although ten different types of fractions were used by the participants, the most commonly used were halves, thirds and quarters. The second preferred type of expression is ratios. From the nine different types of ratios employed (ranging from N in 10 to N in 1000), the most common were N in 10 and N in 100. It is surprising that 7.5% of the expressions chosen were percentages, even though participants were asked to simplify for people who do not understand percentages. We are unsure whether they ignored the instructions, did not agree with them, or just did not find another way of simplifying the expression. However, the use of percentages is not significant with respect to the use of non-numeric expressions.

Whole number (cardinal) percentages are the preferred simplification for *people who do not understand decimals*. This reinforces the idea that they are easier to understand than the original number, while at the same time being the closest to the original value and mathematical form. Frequencies of use of fractions and ratios are very similar and are not significantly different. Non-numeric simplifications were seldom used, in contrast to the first part of the survey; in fact, they occurred only for the peripheral points on the proportion scale, e.g., *almost everyone* or *a little*.

Fractions and ratios are similarly used when simplifying for *people with poor numeracy*. The frequencies of non-numeric expressions and percentages are similar to the ones in the first part of the survey.

In order to test hypothesis H1 (round or common values are preferred to precise ones), we carried out a series of two sample t -tests on common and uncommon fractions and ratios. The results showed that there was significant difference between the use of common and uncommon fractions in the three parts of the survey and the whole survey (no percentages: $p < .001$, no decimals: $p = .07$, free simplification: $p < .0001$, whole: $p < .0001$). However, in the case of ratios there was no significant difference except in the case of free simplification (no percentages: $p = .48$, no decimals: $p = .36$, free simplification: $p = .006$, whole: $p = .14$).

As can be seen in the results, the use of different types of fractions seems to depend on the value being simplified, with quarters, thirds and halves (common fractions) preferred in the central range from 20% to 80%, and greater variety (and rarer use of fractions) at the peripheral. These phenomena can also be observed in non-numeric expressions. This was our hypothesis H2, and in order to test it we performed a series of two sample t -tests on the use of fractions, ratios, percentages and non-numeric in central and peripheral values. The results showed that the use of the four strategies was significantly different for central and peripheral values of the proportions (fractions: $p < .0001$, ratios: $p = .03$, percentages: $p < .0001$, non-numeric: $p < .0001$). The only exception was the use of ratios in the first part of the survey (simplification for people who do not understand percentages), with a p -value of 0.14.

5 Discussion

When asked to simplify for people who do not understand percentages, or for people with poor numeracy, the participants preferred fractions, followed by ratios; when asked to simplify for people who do not understand decimals, they preferred whole-number percentages. Responses show that fractions are considered as the simplest mathematical form, followed by ratios, but this did not mean that fractions were preferred to ratios in every case: the value of the original proportion also influenced choices, with fractions heavily preferred for central values (roughly in the range 0.2 to 0.8), and ratios or non-numeric preferred for peripheral values (below 0.2 or above 0.8), always depending on the kind of simplification being performed.

As some participants commented, not only are percentages mathematically sophisticated forms, but they may be used in sophisticated ways in a text, often for example describing rising and falling values, for which increases or decreases can themselves be described in percentage terms. Such complex relationships are likely to pose problems for people with poor numeracy even if a suitable strategy can be found for simplifying the individual percentages. Another danger is that simplifying several related percentages might obscure the relationship between them. One obvious case would be to render two different values identical, e.g. by simplifying both 48% and 52% to one-half. Another would be to replace two percentages by apparently simpler raw data, for two values with different totals, thus making it harder to see which of the two proportions is larger. In some of the examples with more than one numerical expression being compared, some of the evaluators reported a tendency to phrase them both according to a comparable base - e.g., both in terms of tenths, rather than one as a fifth and one as a third. Thus we should consider the role of context (the set of numerical expressions in a given sentence as a whole, and the meaning of the text) in establishing what simplifications must be used.

6 Conclusions

Through a survey administered to experts on numeracy, we have collected a wide range of examples of appropriate simplifications of percentage expressions. Our aim is to use this data to guide the development of a system for automatically simplifying percentages in texts. With the knowledge acquired from our study we will improve our algorithm to simplify numerical expressions. Our initial hypothesis was that in choosing suitable simplifications, our experts would favor certain mathematical forms - those corresponding to simpler mathematical concepts taught earlier in the curriculum. As expected, the results supported a ranking in which fractions were the simplest form, followed by ratios, whole-number percentages and decimal percentages. However, it did not follow that for *any* proportion value a fraction was the most appropriate simplification, because other forms (e.g., non-numeric expressions) were preferred for peripheral values (near to 0% or 100%) that would have required unfamiliar fractions such as one-hundredth. The value of the original proportion also influenced choices, depending on its correspondence with central or peripheral values. Our results also show that the experts use different options for each simplification strategy.

We have also collected a parallel corpus of numerical expressions (original and simplified version). This corpus will be shared with other researches so it can be used to different applications to improve the readability text. This could be a very useful resource because simplification of percentages remains an interesting and non-trivial problem.

References

1. Nations, U., Standard Rules on the Equalization of Opportunities for Persons with Disabilities. Technical report (1994)
2. Williams, J., Clemens, S., Oleinikova, K., Tarvin, K.: The Skills for Life survey: A national needs and impact survey of literacy, numeracy and ICT skills. Technical Report Research Report 490, Department for Education and Skills (2003)
3. Krifka, M.: Be brief and vague! And how bidirectional optimality theory allows for Verbosity and Precision. In: *Sounds and Systems: Studies in Structure and Change: A Festschrift for Theo Vennemann*. Trends in Linguistics, vol. 141, pp. 439–458. Mouton de Gruyter, Berlin (2002)
4. Williams, S., Power, R.: Precision and mathematical form in first and subsequent mentions of numerical facts and their relation to document structure. In: *Proc. of the 12th European Workshop on Natural Language Generation*, Athens (2009)
5. Grice, H.P.: Logic and Conversation. In: Cole, P., Morgan, J.L. (eds.) *Syntax and Semantics*. Speech Acts, vol. 3, pp. 41–58. Academic Press, San Diego (1975)
6. Chandrasekar, R., Doran, C., Srinivas, B.: Motivations and Methods for Text Simplification. In: *COLING*, pp. 1041–1044 (1996)
7. Siddharthan, A.: Resolving Attachment and Clause Boundary Ambiguities for Simplifying Relative Clause Constructs. In: *Proceedings of the Student Research Workshop, 40th Meeting of the Association for Computational Linguistics* (2002)
8. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of English newspaper text to assist aphasic readers. In: *AAAI 1998 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Wisconsin (1998)
9. Williams, S., Reiter, E.: Generating readable texts for readers with low basic skills. In: *Proceeding of the 10th European Workshop on Natural Language Generation*, Aberdeen, Scotland, pp. 140–147 (2005)
10. Petersen, S.E., Ostendorf, M.: Text Simplification for Language Learners: A Corpus Analysis. In: *Speech and Language Technology for Education (SLaTE)* (2007)

Blind People and Mobile Keypads: Accounting for Individual Differences

Tiago Guerreiro, João Oliveira, João Benedito, Hugo Nicolau,
Joaquim Jorge, and Daniel Gonçalves

INESC-ID / Technical University of Lisbon, Portugal
{tjvg,hman,jaj,djvg}@immi.inesc-id.pt,
{jmgdo,jpmlb}@ist.utl.pt

Abstract. No two persons are alike. We usually ignore this diversity as we have the capability to adapt and, without noticing, become experts in interfaces that were probably misadjusted to begin with. This adaptation is not always at the user's reach. One neglected group is the blind. Age of blindness onset, age, cognitive, and sensory abilities are some characteristics that diverge between users. Regardless, all are presented with the same methods ignoring their capabilities and needs. Interaction with mobile devices is highly visually demanding which widens the gap between blind people. Herein, we present studies performed with 13 blind people consisting on key acquisition tasks with 10 mobile devices. Results show that different capability levels have significant impact on user performance and that this impact is related with the device and its demands. It is paramount to understand mobile interaction demands and relate them with the users' capabilities, towards inclusive design.

Keywords: Individual Differences, Mobile Accessibility, Blind, Mobile Device, User Assessments.

1 Introduction

Mobile devices have become indispensable tools in our daily lives and are now used by *everyone* in several different situations. However, every human is different and so is every situation. This diversity has not been given enough attention in mobile user interface design decreasing the user's effectiveness or even hindering the ability to interact. Particularly, disabled target groups, characterized by specific individual differences, can greatly benefit from an effective mobile user interface [1]. However, alternative user interfaces are likely to be misaligned with the users and their identities. These adapted mobile user interfaces are stereotypical disregarding that the relation between the user and the device depends on particular characteristics and not on a common idea. In general, mobile interaction has not evolved to meet the users' requirements.

We focus our attention on blind people. The absence of such an important and integrating sense as vision, in the presence of so demanding interfaces as are the ones in current mobile devices, justifies it. Besides personality differences, two blind users

are likely to have totally different stories to what blindness, and its implications, is concerned. The cause of the impairment, age of onset, time with impairment, age, simultaneous impairments, cognitive or sensorial abilities, are some examples of the characteristics that may diverge between users. Some of these may implicate others, some may be collateral damages, and others can overcome some. A young '*recent-blind*' is different from an older one. While the former is likely to have all his other senses immaculate, the latter may have some other age-related impairments. However, he is also likely to have developed sensory compensation mechanisms. How are they different and how will those differences affect their functional ability?

What is indisputable is that the enormous diversity found among this particular group of users turns the "stereotypical blind" idea inadequate. As an example, age, its related degenerations, but also how it influences sensory compensation and augments individual differences, is an example of a characteristic that must be considered when discussing a particular blind person. As blindness age of onset can determine how one will face the challenges imposed by any daily task. These characteristics have a huge impact on the blind users' lives and how they are able to deal with technology. Regardless, all are presented with the same methods and opportunities ignoring their capabilities and needs.

Interaction with mobile devices is highly visually demanding which increases the difficulties. Even mobile assistive technologies for the blind have a narrow and stereotypical perspective over the difficulties faced by their users. A blind user is presented with screen reading software to overcome the inability to see onscreen information. However, these solutions go only half-way. In the absence of sight other aptitudes/limitations stand up. To empower these users, a deeper understanding of their capabilities and how they relate with technology is mandatory. As such, we performed a study with the target population consisting on key acquisition tasks with different mobile keypads, in order to relate individual differences with devices' demands.

1.1 The Blind

We focus on a particular target group: blind people. This can be explained both by the high visual demands imposed by current mobile device interfaces (which are increasing, e.g., touch screens) and the diversity within the population. In this section, we offer an overview of the population and related concepts, valuable to understanding the remaining of this paper.

1.1.1 Causes

Blindness is due to a variety of causes. Age-related blindness is increasing throughout the world, as is blindness due to uncontrolled diabetes. Diabetes is responsible for 8% of legal blindness, according to the American Diabetes Association, *making it the leading cause of new cases of blindness in adults 20-74 years of age*. This is significant since diabetic retinopathy is often accompanied by peripheral neuropathy which also impairs the sense of touch [2].

1.1.2 Worldwide Statistics

The American Foundation for the Blind estimates that there are 10 million blind or visually impaired people in the United States. In a survey realized in 1994-1995, 1.3 million Americans (0.5%) reported being legally blind. Of this number, only 10% were totally blind and another 10% had only light perception. The remaining 80% had some useful vision. Few statistics appear to be available about the age of onset of blindness. It is reported that *only eight percent of visually impaired people are born with any impairment* [3]. Worldwide, an estimated 180 millions are visually impaired, of which 40-45 millions are blind [4].

The prevalence of blindness is much higher for the elderly [2]. It is estimated that 1.1% of the elderly (65 and over) are legally blind compared to 0.055% of the young (20 and under) [5]. About 82% of all people who are visually impaired are age 50 and older (although they represent only 19% of the world's population). The attitude towards blindness as well as space representation may be affected by the age of onset of blindness [2]. It is also reported that more than 50 percent of individuals with visual impairments also have one or more other impairments [6]. It is worth mentioning that blindness is expected to increase in the following years [5].

1.1.3 Individual Differences among Blind People

Previous studies [7] have shown that individual differences among blind people are likely to have a wider impact on their abilities to interact with mobile devices than among sighted people. Tactile sensitivity, spatial ability, short-term memory, blindness onset age and age are mentioned as deciding characteristics for a blind user mobile performance.

The capability-demand theory builds on the concepts of user capability and product demand and aims to analyze user-product compatibility, i.e. *an assessment and comparison of the sensory, cognitive and motor demand made by a product in relation to the ability levels of the expected user population* [8]. We embrace the capability-demand theory and aim to assert relations between users and devices, and ultimately aim at a match between individual capabilities and mobile interaction demands. This way, we will provide both the tools for mobile designers, showing which designs are most effective and inclusive, and for blind users, identifying for each one, the most appropriate interfaces.

2 Related Work

Mobile interaction is still in its early stages when compared with interaction with desktop computers that has been subject of attention for several decades. Although mobile computing is an active research theme, mobile HCI was not until recently an important subject. In particular, only few researchers have leaned over the multitude of individuals, scenarios and situations faced by mobile devices.

The related work presented here is twofold. Firstly, we survey mobile user interfaces for blind users and present the actual research state considering alternative interfaces for the target population. Secondly, we present projects that have focused on

individual differences, taking into consideration some kind of particular characteristics instead of a global "average user" model.

2.1 Mobile Interaction for the Blind

A blind person, or even one with low vision, faces several limitations when interacting with mobile devices [9,10,11]. Looking at these, whether keypad or touch screen-based, the interaction mechanisms are convoluted to deal with the limited input area and overall device small size. Furthermore, the mechanisms found to overcome the lack of space (when compared to desktop computers) resort to an intensive visual-based dialogue with the mobile device user. As an example, several text-entry systems are based on multi-tap approaches where the user is able to see both the relation between keys and letters (visual feedback from the physical or virtual keys), and the evolution of the process on the display. A user with severe visual limitations is unable to receive this information and thus his ability to interact with these devices is highly limited.

An attempt to attenuate difficulties arising from disabilities is being made through assistive technologies, designing new solutions based on alternative interaction methods such as haptic interfaces, screen readers or multimodal information feedback [12]. In the case of mobile phones, accessibility solutions consist of devices with native features to make them more accessible to a given population, the so called accessible phones, or mobile phones that allow third-party software installation like screen readers or screen magnifiers.

Special mobile devices were developed to overcome the barriers arising from visual impairments. As examples are the *Brailino* or the *Alva Mobile Phone Organizer*, among many others very similar between each other [13]. These devices, which typically work as a Personal Digital Assistant, or as a docking station, use a Braille keyboard for text input, a Braille screen for output information, and provide functionalities like the ones provided in regular mobile phones. Yet, they all share the same flaws: their cost is prohibitive and they are not as portable as a mobile phone is, being too big and heavy. Even though their cost and size has decreased, they are still not as practical as common mobile devices and have the inconvenient to "look disabled". Another factor, and one of the most important, is that although Braille is the recognized alternative language for blind people, a reduced percentage is actually Braille-knowledgeable¹.

Nowadays, a common mobile solution for blind users resorts to the usage of a screen reader, replacing the visual feedback by its auditory representation (e.g., *Mobile Speak*² or *Nuance Talks*³) [14]. This approach supposedly enables the users to access the same applications a fully-sighted user accesses. The ability to use a "non-disabled" device with the same characteristics (technical, social and economical) is a great advantage of this type of solutions. However, the offered feedback is restricted

¹ American Foundation for the Blind: Programs and Policy Research, "Estimated Number of Adult Braille Readers in the United States", International Braille Research Center (IBRC), <http://www.braille.org/papers/jvib0696/vb960329.htm>

² <http://www.codefactory.es/en/products.asp?id=316>

³ <http://www.nuance.com/for-individuals/by-solution/talks-zooms/index.htm>

to the output, as no information is obtained on key/function relation. Moreover, the information on the screen is prepared for visual feedback and not to be read. As an example, considering text-entry, screen reader approaches force the user to try to find the desired letter in the keypad, committing several errors in the process, and possibly leading to situations where he/she simply quits trying. A person that acquires blindness in an advanced stage of life, along with the reduction of other capabilities such as tactile sensitivity, is likely to face difficulties in the first contact with this approach, rejecting it before gaining the experience that enables its use [9]. In contrast to traditional interfaces, designed for the "average user", simple screen reading approaches are designed for the "stereotypical blind", one that has improved tactile and auditory senses along with good mental health and motivation.

Guerreiro et al. [15] proposed two non-visual texting interfaces for keypad-based devices - NavTap and BrailleTap – that take advantage of blind users' capabilities. NavTap was designed for those who are not able to learn traditional MultiTap methods, allowing them to easily navigate through the alphabet using only four keys, thus eliminating the need to memorize mappings between keys and letters. Similarly, by transforming the traditional keypad layout into a more familiar interface, BrailleTap was designed for those who master the Braille alphabet, allowing them to enter text on common devices. More recently, with the emergence of touch screen devices, several approaches have been proposed featuring directional gestures for menu navigation [16] and text-entry tasks [17,18].

In general, there has been an effort to provide visually impaired users with technologies that are able to surpass the target group's problems and inabilities. However, the majority of the approaches are designed taking into consideration a "stereotypical" image of the blind user. Whether considering Braille interfaces or other touch-demanding interfaces, they are created without a real knowledge of the users and their needs. If blind-targeted interfaces were designed following an user-centered design approach, nowadays they would not be merely based on the Braille alphabet (as the overall knowledge is minimum) nor on the users' tactile capabilities (as a great majority of the blind population are older adults with no compensatory mechanisms and low tactile capabilities). However, there are interfaces characteristics that have shown to be advantageous for particular sets of users. Further research needs to be performed to assess the individual differences within the target population and understand the interface demands and match for particular blind users.

2.2 Accounting for Individual Differences

While desktop computers already offer a multitude of personalization capacities, whether considering individual differences or just taste, mobile devices are still restricted to a limited set of personalization options, which are majorly aesthetic or related to the users' personality. A better understanding of the individual differences that characterize the users in the mobile interaction context is required. Only with that knowledge will we be able to "prescribe" adequate devices and interfaces that empower the users.

While it is important to understand that a mobile device user is different from the next one and that those differences should be considered to improve device accessibility, it is also important to understand that, even for a single user, his capacities and

needs are likely to diverge across time (dynamic diversity) [19]. Gregor and Newell state that most computer systems are designed for a typical younger user with static abilities over time. However, even when user-centered paradigms are employed, the designers look typically at concerns such as representative user groups, without regard for the fact that the user is not a static entity. This does not take into account the wide diversity of abilities among users and it also ignores the fact that these abilities are dynamic over time. The authors propose a new paradigm, designing for Dynamic Diversity, based on a user-sensitive inclusive design methodology [19]. *The use of the term "inclusive" rather than "universal" reflects the view that "inclusivity" is a more achievable, and in many situations, appropriate goal than "universal design" or "design for all". "Sensitive" replaces "centered" to underline the extra levels of difficulty involved when the range of functionality and characteristics of the user groups can be so great that it is impossible in any meaningful way to produce a small representative sample of the user group nor often to design a product which truly is accessible by all potential users [19].*

In order to lessen the difficulties that many people feel when interacting with mobile devices, inclusive design when developing these products is crucial. To make this a reality, it is necessary to attend to the many different characteristics of the user. Persad et al. [8] propose an analytical evaluation framework based on the Capability-Demand theory, where user capabilities at sensory, cognitive and motor levels, are matched with product demands.

When studying interfaces for blind people, a capability that should not be ignored is tactile sensibility. Besides being crucial to capture information at the expense of vision, approximately 82% of all people who are blind are aged 50 or more [20] and as diabetes is one the main causes of blindness, changes in this sensorial capability are fairly common and should be accounted for. In [21,22] several physical requirements were identified in order for mobile devices to be accessible with limited sensibility. Despite the fact that these studies acknowledged key requirements, these characteristics were not quantified nor related with the different users' abilities.

Cognitive capabilities such as short-term memory, attention and spatial ability should also be meaningful when developing interfaces for the blind. Mobile interaction requires a cognitive effort that, for someone lacking sight, is much more demanding. Although there are studies that relate cognitive ability with mobile device usage for sighted older adults [23], there is an enormous gap in terms of studies relating cognitive ability and mobile phone interaction of a visually impaired person.

3 Assessing the Impact of Individual Differences

Mobile accessibility solutions for blind users are commonly presented as audio replacements for onscreen information. However, there is more to it. Not only the interaction is much more demanding in the absence of sight, but also blind people often present differences that make them far in capability from the stereotyped blind.

3.1 Research Goals

In this study, we intend to understand the impact of individual differences among the blind when interacting with mobile device keypads as well as how these differences are revealed when confronted with different device demands. Particularly, we want to answer the following research questions: 1) Does tactile sensitivity affect performance on a simple key acquisition task?; 2) Does cognitive ability affect performance on a simple key acquisition task?; 3) Which keyboard demands have most impact on user performance?; 4) Are individual differences worth considering on mobile interface design?

3.2 Procedure

To evaluate the match between users and interfaces by using various measures of compatibility, we performed studies with the target population. They were threefold. The first two consisted in measuring user capabilities, more specifically, tactile sensitivity and cognitive ability. In the third and final, we conducted an experimental task to assess user performance in a simple key acquisition task. Details on each of these studies and the results obtained considering the different capabilities and product demands are depicted in the following sections.

3.3 Tactile Sensitivity

To assess the participants' tactile capabilities, two different components of tactile sensitivity were measured. The first, pressure sensitivity, was determined using the Semmes-Weinstein monofilament test [24] (Figure 1). In this test, there are several nylon filaments with different levels of resistance, bending when the maximum pressure they support is applied. This way, if a user can sense a point of pressure, his pressure sensitivity is equal to the force applied by the filament.

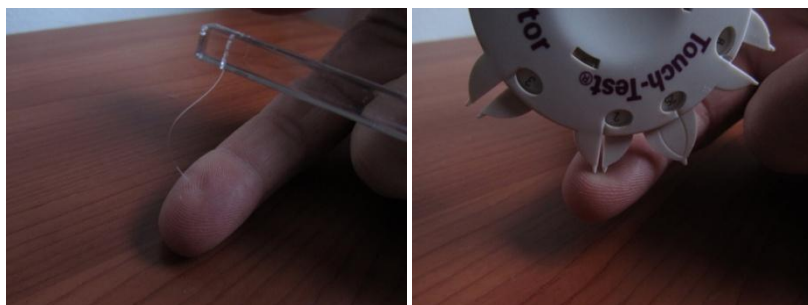


Fig. 1. Tactile assessment: Semmes-Weinstein test (left) and Disk-Criminator (right)

Five monofilaments of 2.83, 3.61, 4.31, 4.56 and 6.65 Newton were used, starting the stimuli with the one of 2.83, the least resistant one. Pressure was applied in the thumb, index and middle fingers, those generally used when interacting with mobile

devices, and in random order, so we could prevent arbitrary identification of a stimulus by the person being tested. The process is repeated with the filament with the next resistance level, until all filaments are tested or the participant correctly identifies the stimulus made. Different levels can be found for different fingers.

The other tactile sensitivity component measured was spatial acuity, using the Disk-Criminator [25] (Figure 1). This instrument measures a person's capability to distinguish one or two points of pressure on the skin surface. The Disk-Criminator is generally an orthogonal plastic instrument that has in each side a pair of metal filaments with relative distances ranging from 1 to 25mm. When the person being tested identifies a stimulus as being two points, her spatial acuity discrimination is equal to the distance between the filaments.

The distance between the filaments of the Disk-Criminator tested ranged from 2 to 15 mm, with 1mm increments. Each of these filament pairs was, as with the pressure sensitivity, applied randomly in the same three fingers. There were made 10 stimuli per finger, randomly, alternating between a pair of filaments and a unique filament. The participant had to indicate when he/she felt one or two points of pressure. When he/she was able to correctly identify 7 out of 10 stimuli, his/her level of spatial acuity was registered as the distance between filaments.

3.4 Cognitive Assessment

The cognitive evaluation focused two components of the cognitive ability, a verbal and a non-verbal. The verbal component was evaluated in terms of working memory: a short-term memory and main responsible for the control of attention [26]. The non-verbal component, which consists of abilities independent of mother language or culture, was evaluated in terms of spatial ability: the ability to create and manipulate mental images, as well as maintain orientation relatively to other objects [27].

To evaluate working memory, the subtest Digit Span of the revised Wechsler Adult Intelligence Scale (WAIS-R) was used [28]. In the first part of this test, the participant must repeat increasingly long series of digits presented orally, and on the second, repeat other sets of numbers but backwards. The last number of digits of a series properly repeated allows calculation of a grade to the participant's working memory and, subsequently, to the user's verbal intelligence quotient (Verbal IQ).

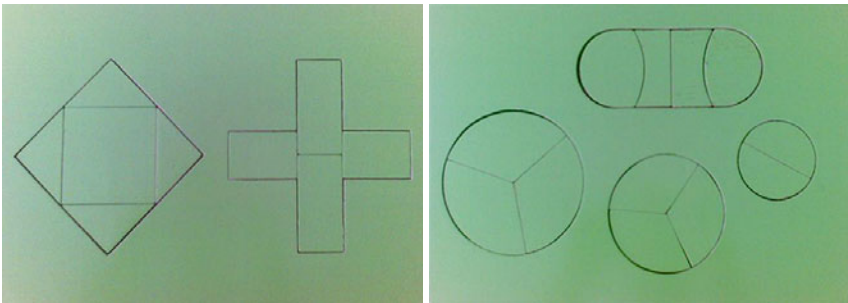


Fig. 2. Spatial ability test: Planche a Deux Formes (left) and Planche du Casuiste (right)

Spatial ability was measured using the combined grades of the tests *Planche a Deux Formes* and *Planche du Casuiste*. These two tests are part of a cognitive battery for vocational guidance [29]. The goal of these tests is to complete, as fast as possible, a puzzle of geometrical pieces (Figure 2).

A K-means algorithm was used to group the *Spatial Ability* and *Verbal IQ* values into levels according to their different measures. The different values were grouped into three levels: *Inferior*, *Average*, *Superior* (Table 1).

Table 1. The studied cognitive ability characteristics and the group levels formed with a K-Means algorithm. Each group level's measure corresponds to the cluster's centre value.

Attribute	Group Levels			Measure
	Inferior	Average	Superior	
Spatial Ability	1.90	5.20	9.03	Spatial ability score (1-20).
Verbal IQ	72	91	129	Verbal IQ score.

3.5 Experimental Task

While we acknowledge that the challenges imposed to blind users are spread among different interaction layers, we focused our attention in the first barrier they encounter, the physical one. Therefore, we chose the simplest task when using a mobile device, so we could isolate its demands and understand their effect on users' performance. The task presented in this study consisted of simple key acquisitions with 10 different mobile devices (Figure 3). All selected phones featured a keypad, since they are still the most used among the blind. None of them had any sort of accessibility aid. Devices were categorized according to their keypads' physical attributes. A K-means algorithm was used to group these attributes into levels according to their different measures. *Key Spacing* and *Key Height* values were grouped into three levels: *Smaller*, *Average* and *Bigger*; while *Key Size* and *Key Label* into two: *Smaller* and *Bigger*. The values of the last attribute *Key Material* were grouped into two categories: *Good*, if the materials on the phone and keyboard are distinct, and *Bad* if not. Table 2 shows the keyboard attributes and their groups, including the value of the clusters' center for each level.



Fig. 3. Mobile phones tested (from top left to bottom right: Samsung C108, Nokia 1110, Motorola C650, Nokia 6150, Siemens MC 60, Nokia 3310, Nokia 638, Sendo, Samsung, HTC S310). A test with a user in three of the ten mobile devices (right).

Our goal in this experimental task was to evaluate performance, in terms of efficiency (time) and effectiveness (number of errors), in low level tasks. Each participant was asked, randomly, to input a key (0 to 9, “*” and “#”) with the keypad. All numbers were issued two times making a total of 24 keys by phone (Figure 3). After each acquisition, the participant had to place the operating hand resting on the table.

All of the mobile devices were shown, before the evaluation, to the volunteers, so they could experiment and understand the layout. All doubts were cleared by the researchers.

The time taken to press each key was measured (*Average Time*) as well as the number of errors (*Task Errors*), with an error being considered when the participant pressed the wrong key, pressed more than a key, or tried to press a key but failed.

Table 2. The keyboards’ attributes and the group levels formed with a K-Means algorithm. Each group level’s measure corresponds to the cluster’s center value.

Attribute		Group Levels			Measure
		Smaller	Average	Bigger	
Key Spacing	Horizontal	0.7 mm	2.2 mm	4.6 mm	Distance between the closest edges of the keys, horizontally and vertically (mm).
	Vertical	0.8 mm	2.2 mm	3.3 mm	
Key Height		0 mm	0.5 mm	1.1 mm	Distance between the base and the top of the key (mm).
Key Size	Horizontal	8.3 mm		11.7 mm	Largest side or diameter of the polygon, horizontally and vertically (mm).
	Vertical	4.8 mm		6.1 mm	
Key Label		0.1 mm		0.4 mm	Height of the label(s) on the number 5 key (mm).
Key Material					If the key material is contrasting with the mobile phone's base material.

3.6 Participants

The participant group was composed by 13 students (8 female, 5 male) from a formation centre for visually impaired people, where all the evaluations took place. With ages ranging from 25 to 61 years old (averaging 50), all of the volunteers were blind (at most light perception).

All of the participants used mobile phones on a daily basis with the help of screen readers. However, even when using screen readers, 3 participants stated that they have difficulty sending text messages, restricting use to placing and receiving calls. Their characterization, including the results of the tactile sensitivity and cognitive assessments, is displayed in Table 3.

3.7 Design and Analysis

This study features two dependent variables (*Average Time* and *Task Errors*) and several factors related to device attributes (*Key Size*, *Key Spacing*, *Key Height*, *Key Material* and *Key Label*) as well as participants’ individual differences (*Blindness Onset*, *Time with Impairment*, *Education*, *Spatial Acuity*, *Pressure Sensitivity*, *Spatial Ability*, *Verbal IQ*).

Table 3. Participant's characterization. The lower the spatial acuity and pressure sensitivity values, the better the tactile sensitivity. The opposite for verbal IQ and spatial ability.

User	Gender	Age	Time with Impairment	Education	Spatial Acuity	Pressure Sensitivity	Verbal IQ	Spatial Ability
P01	Female	48	2 Years	4 th Grade	3	4.31	66	1,75
P02	Female	58	54 Years	6 th Grade	4	3.61	69	1,75
P03	Female	53	21 Years	7 th Grade	3	4.31	84	3,25
P04	Female	48	31 Years	9 th Grade	2	4.31	78	5,5
P05	Female	39	5 Years	9 th Grade	2	4.31	79	7,75
P06	Male	61	58 Years	6 th Grade	2	4.31	104	4,75
P07	Male	25	6 Years	12 th Grade	2	2.83	64	5,5
P08	Female	53	33 Years	4 th Grade	4	4.31	78	1
P09	Male	51	16 Years	9 th Grade	3	4.31	84	6,25
P10	Male	59	59 Years	9 th Grade	2	4.31	134	4
P11	Female	41	21 Years	12 th Grade	2	3.61	123	10,85
P12	Male	58	24 Years	9 th Grade	2	4.31	86	8,5
P13	Female	55	55 Years	10 th Grade	2	3.61	98	1,75

Shapiro-Wilk test was used to examine the normality of the data. Since the two dependent variables did not exhibit a normal distribution, Kruskal-Wallis test was used to assess significant differences of the different independent variables. When statistically analysing multiple factors, a two-way ANOVA was used, admitting data normality. Correlation between *Blindness Onset and Time with Impairment* with the dependent variables was measured using Pearson's Correlation Coefficients.

4 Results

Our goal is to understand the impact of each characteristic and how these differences are revealed when subject to different demands. First, we focus on each user attribute and present the impact on overall performance. Then, we look into device demand variations and observe how they affect participants' effectiveness and efficiency. Hence, we will be able not only to call researchers and manufactures attention to individual differences but also to point out the demands that should be considered to promote inclusive design.

4.1 Individual Differences

Blindness Onset age had a positive low degree correlation with *Task Errors* ($r=0.22$, $p<.05$). It also had significant moderate correlation with *Average Time* ($r=0.29$, $p<.01$). The greater the age of onset, the less time and number of errors the users made. A relation between *Time with Impairment* and *Average Time* was also found. The bigger the *Time with Impairment*, the less *Average Time* the users took, as a moderate negative correlation was found ($r=0.37$, $p<.01$).

The user's *Education* had an impact on the *Task Errors* ($X^2(2)=8.34$, $p<.01$). More educated user's made significantly fewer errors on the task proposed.

In terms of tactile sensitivity, *Spatial Acuity* had a significant effect on *Average Time* ($X^2(1)=4.204, p<.05$). Participants with better spatial acuity (2mm) were faster than the remaining (Figure 4).

Regarding cognitive ability, *Spatial Ability* did not impact the efficiency or the effectiveness of the volunteers. On the other hand, the *Verbal IQ* relation with *Average Time* and *Task Errors* had a significant effect ($X^2(2)=21.92, p<.01$ and $X^2(2)=6.33, p<.05$). The higher the *Verbal IQ*, less time the users took to press a button (Figure 4), while also committing fewer errors (Figure 5). Although *Spatial Ability* has not shown significant effects, the two cognitive abilities together had a significant influence on the outcome of the user’s performance ($F_{3,122}=5.63, p<.01$).

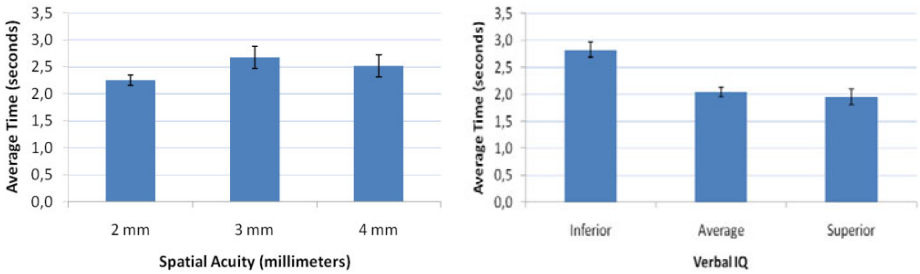


Fig. 4. *Average Time* for *Spatial Acuity* (left) and *Verbal IQ* (right). Error bars denote 95% CI.

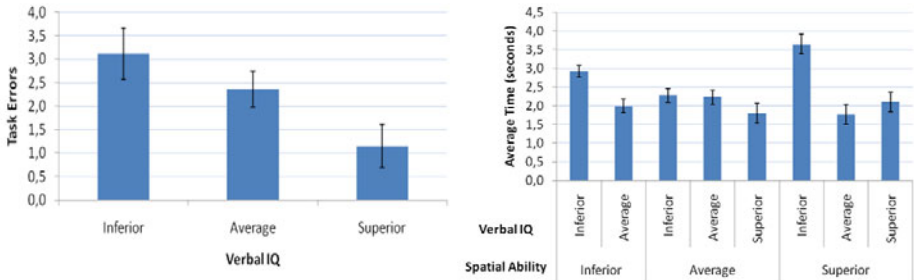


Fig. 5. *Task Errors (average)* for *Verbal IQ* (left) and *Average Time* for *Spatial Ability* and *Verbal IQ* (right). Error bars denote 95% CI.

Looking at the relation of cognitive ability with tactile sensibility, a multiple factor analysis shows that the *Verbal IQ* and *Spatial Acuity* affected significantly the efficiency and effectiveness of the participants ($F_{1,124}=4.20, p<.05$ and $F_{1,124}=8.85, p<.01$ respectively). Figure 6 shows that for participants with the same *Spatial Acuity*, performance improved with *Verbal IQ* increase.

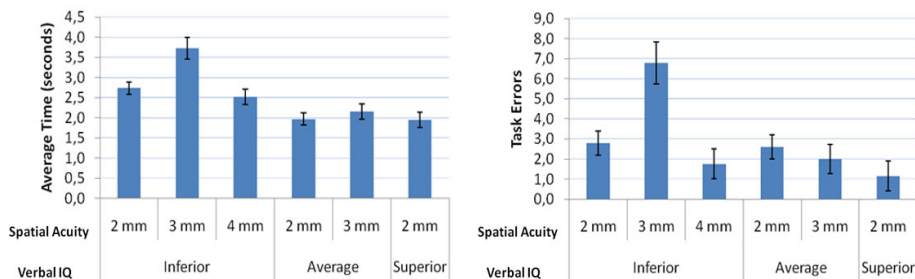


Fig. 6. Average Time for each Verbal IQ and Spatial Acuity (left) and Task Errors (average) for each Verbal IQ and Spatial Acuity (right). Error bars denote 95% CI.

4.2 Mobile Devices

The analysis of the participants' performance on each mobile phone shows that the number of errors and time varies across all devices (Figure 7). The analysis of variance shows that the devices' attributes influence significantly the users' efficiency ($X^2(9)=39.81, p < .01$) and effectiveness ($X^2(9)=46.72, p < .01$).

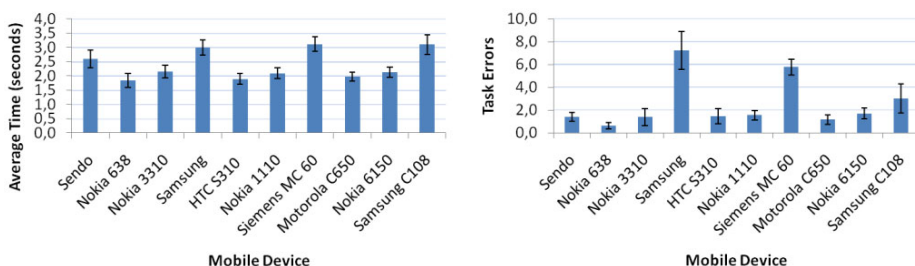


Fig. 7. Average Time for each Mobile Device (left) and Task Errors (average) for each Mobile Device (right). Error bars denote 95% CI.

If we look at each attribute by itself, all of them, with the exception of *Key Material*, had a significant effect on *Task Errors* ($X^2(1)=7.68, p < .01$ for *Key Size*, $X^2(2)=13.26, p < .01$ for *Key Height*, $X^2(2)=10.52, p < .01$ for *Key Spacing* and $X^2(1)=7.28, p < .01$ for *Key Label*). As we can see in Figure 8, the bigger and more pronounced the attribute was, fewer errors participants made.

The relation of the attribute *Key Height* with *Average Time* was also significant ($X^2(2)=16.12, p < .01$), as we can see on Figure 9.

A multiple factor analysis between the different characteristics of the mobile devices revealed a couple of significant pairs. The pair *Key Size* and *Key Height* had a significant effect on *Average Time* and *Task Errors* of the participants ($F_{2,124}=14.47, p < .01$ and $F_{2,124}=2.84, p < .01$ respectively). We can verify from Figures 9 and 10 that participants made by far most mistakes and took the longest time when interacting

with devices with a small *Key Spacing* and an average *Key Size*. It is also interesting to note that a better performance resulted from a higher *Key Spacing* between keys in conjunction with average *Key Size*, than with the opposite.

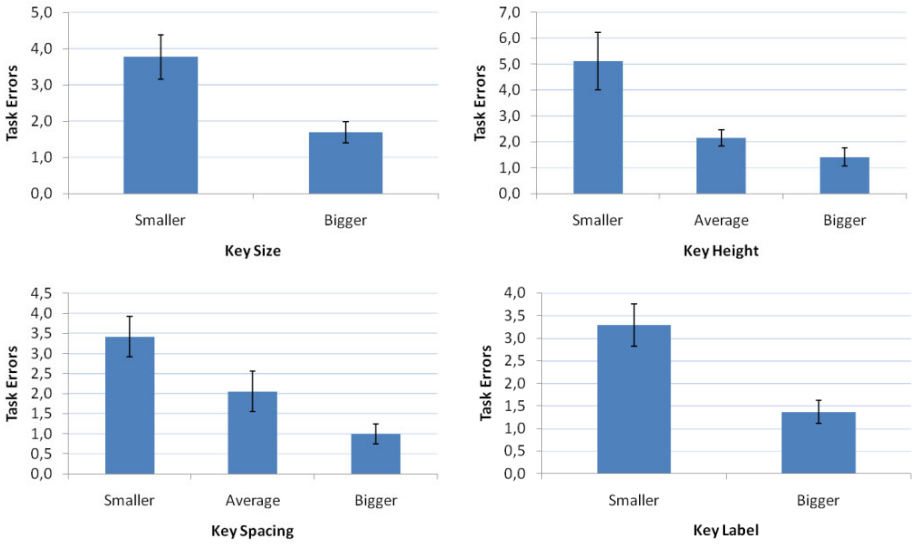


Fig. 8. Task Errors (average) for each Key Size (top left), Key Height (top right), Key Spacing (bottom left) and Key Label (bottom right). Error bars denote 95% CI.

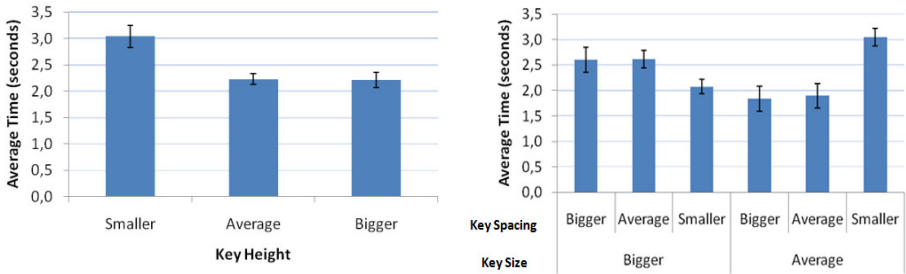


Fig. 9. Average Time for Key Height (left) and for Key Size and Spacing (right). Error bars denote 95% CI.

Key Height and *Key Size* had a significant effect on *Task Errors* ($F_{2,124}=3.28, p<.05$). Figure 10 showcases this relation, where we can observe that the devices with an average *Key Size* but with a small *Key Height* resulted in poor effectiveness. It is also apparent that independently of *Key Size*, the smaller the *Key Height* was, the more errors participants made.

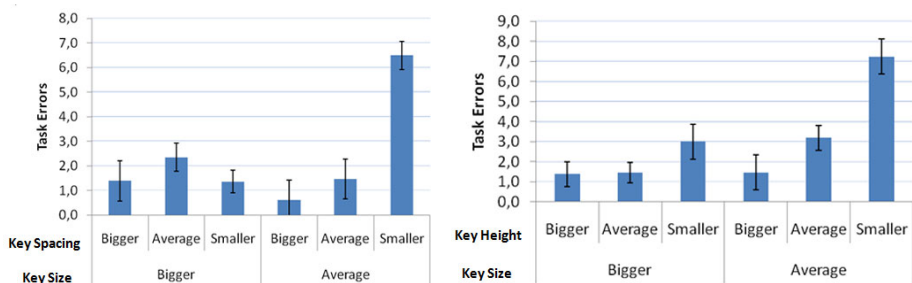


Fig. 10. Task Errors (average) for each Key Size and Key Spacing (left) and Task Errors (average) for each Key Size and Key Height (right). Error bars denote 95% CI.

Other pair with significant impact on users' *Average Time* and *Task Errors* was *Key Label* and *Key Spacing* ($F_{1,125}=17.13$, $p<.01$ and $F_{1,125}=18.70$, $p<.01$ respectively). A small *Key Label* in conjunction with a small *Key Spacing*, resulted in more errors and longer time than any other case, as we can see in Figure 11.

Lastly, *Key Material* and *Key Height* affected significantly ($F_{1,125}=19.40$, $p<.01$) the number of errors made by the participants. Figure 11 shows that keys with a good *Key Material* do not have a clear benefit if they have a small *Key Height*, as the number of errors increased as the height of the keys decreased.

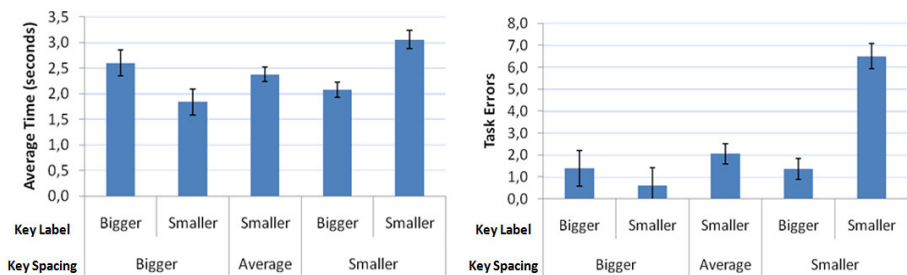


Fig. 11. Average Time for each Key Spacing and Key Label (left) and Task Errors (average) for each Key Spacing and Key Label (right). Error bars denote 95% CI.

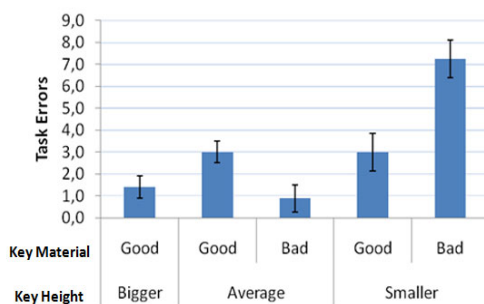


Fig. 12. Task Errors (average) for each Key Height and Key Material. Error bars denote 95% CI.

5 Discussion

Looking back to the aforementioned research questions:

1. *Does tactile sensitivity affect effectiveness/efficiency on a simple key acquisition task?*

Results showed that spatial acuity significantly affects the way the users acquire the keys. Users with less acuity take longer to explore the keypad and recognize the desired keys. This is more visible in tactile-wise demanding devices (low key spacing, size and relief). No significant effect was found on Task Errors. However, this could be explained with the absence of pressure as no time limits were imposed.

2. *Does cognitive ability affect effectiveness/efficiency on a simple key acquisition task?*

Verbal IQ presented a significant effect both on users' effectiveness and efficiency. The evaluations performed are suitable assessments both for attention and short-term memory. Indeed, these are abilities that are likely to change from one individual to another, particularly with age. Regarding spatial ability, no significant effects were found. This was expected as all participants were experienced with the keypad layout.

3. *Which keyboard demands have most impact on user performance?*

All the device characteristics analyzed showed to have influence on user performance. The devices present similar keypad layouts and the task is quite simple. However, differences were still found and, in some cases, dramatic. It is paramount to understand each demand and its limits. Further, this understanding goes beyond accessibility and assistive technologies. Mobile devices are challenging for everyone, every now and then, and inclusive design should be promoted for every user's benefit.

4. *Are individual differences worth considering on mobile interface design?*

Results showed that individual differences have an impact on user performance and hence they should be considered to improve effectiveness. As the tasks demands increase, likely so will the differences between users. These differences are commonly so dramatic that exclusion is reached.

6 Conclusions

Individual differences among the blind have a great impact on the different mobile interaction proficiency levels they attain. General-purpose interfaces and assistive technologies disregard these differences. In this paper, we argue that both the users' capabilities as the device demands should be explored to foster inclusive design. We presented a study with 13 blind people where different capability (particularly, attention and spatial acuity) levels showed to have a significant impact on key acquisition effectiveness and efficiency. It is noticeable that these differences are present even in

the simplest task. Furthermore, mobile device demands variations showed to have a wide impact on the user's ability to interact, giving relevance to informed design.

Next, we will relate low-level user characteristics with mobile interaction modalities demands. In possession of a thorough characterization of how different users relate to different interaction modalities, we will derive a model that allows us to make predictions regarding the performance of particular user/modality pairs.

References

1. Manduchi, R., Coughlan, J.: Portable and mobile systems in assistive technology. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 1078–1080. Springer, Heidelberg (2008)
2. Levesque, V.: Blindness, technology and haptics. Center for Intelligent Machines. Technical Report (2005)
3. Harper, S.: Standardizing electronic travel aid interaction for visually impaired people. UMIST, Technical Report (1998)
4. Leonard, R.: Statistics on Vision Impairment a Resource Manual. Lighthouse International, Technical Report (2001)
5. Hollins, M., Leung, E.: Understanding blindness: An integrative approach. Lawrence Erlbaum, Mahwah (1989)
6. Adams, A.: Electronic Travel Aids: New Directions for Research. Working Group on Mobility Aids for the Visually Impaired and Blind, Committee on Vision, NRC (1986)
7. Guerreiro, T., Jorge, J., Gonçalves, D.: Identifying the individual ingredients for a (in)successful non-visual mobile experience. In: Proceedings of the European Conference on Cognitive Ergonomics, ECCE, Delft, The Netherlands (2010)
8. Persad, U., et al.: Characterising user capabilities to support inclusive design evaluation. *Universal Access in the Information Society* 6(2), 119–135 (2007)
9. Guerreiro, T., et al.: NavTap: a long term study with excluded blind users. In: Assets 2009: Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 99–106. ACM, New York (2009)
10. Kane, S.K., et al.: Freedom to roam: a study of mobile device adoption and accessibility for people with visual and motor disabilities. In: Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 115–122. ACM, New York (2009)
11. Plos, O., Buisine, S.: Universal design for mobile phones: a case study. In: CHI 2006: CHI 2006 Extended Abstracts on Human Factors in Computing Systems, New York, NY, USA, pp. 1229–1234 (2006)
12. Shinohara, K., Tenenberg, J.: A blind person's interactions with technology. *Commun. ACM* 52(8), 58–66 (2009)
13. MacKenzie, I., Tanaka-Ishii, K.: Text entry systems: Mobility, accessibility, universality. Morgan Kaufmann, San Francisco (2007)
14. Burton, D.: You Get to Choose: An Overview of Accessible Cell Phones. *Access Issues* 6(2) (2005)
15. Guerreiro, T., Lagoá, P., Santana, P., Gonçalves, D., Jorge, J.: NavTap and BrailleTap: Non-Visual Texting Interfaces. In: Proceedings of RESNA (2008)

16. Kane, S., Bigham, J., Wobbrock, J.: Slide Rule: Making Mobile Touch Screens Accessible to Blind People using Multi-Touch Interaction Techniques. In: Proceedings of ASSETS, pp. 73–80 (2008)
17. Yfantidis, G., Evreinov, G.: Adaptive Blind Interaction Technique for Touchscreens. *Universal Access in the Information Society* 4(4), 328–337 (2006)
18. Bonner, M., Brudvik, J., Abowd, G., Edwards, W.: No-Look Notes: Accessible Eyes-Free Multitouch Text-Entry. *Pervasive Computers*, 409–426 (2010)
19. Gregor, P., Newell, A.F.: Designing for dynamic diversity: making accessible interfaces for older people. In: WUAUC 2001: Proceedings of the 2001 EC/NSF Workshop on Universal Accessibility of Ubiquitous Computing, pp. 90–92. ACM, New York (2001)
20. Zajicek, M.: Design principles to support older adult, pp. 111–113. Springer, Heidelberg (2004)
21. Kurniawan, S., Mahmud, M., Nugroho, Y. A Study of the Use of Mobile Phones by older Persons. In: Conference on Human Factors in Computing Systems, pp. 989–994 (2006)
22. Kurniawan, S.: Mobile Phone Design for older persons. In: Designing for Seniors: Innovations for Graying Times, pp. 24–25 (2007)
23. Czaja, S., Lee, C.: The impact of aging on access to technology. *Universal Access in the Information Society* 5(4), 341–349 (2007)
24. Tremblay, F., Mireault, A.C., Dessureault, L., Manning, H., Sveistrup, H. *Experimental Brain Research*, 155–164 (2004)
25. Mackinnon, S., Dellon, A.: Two-point discrimination tester. *Journal of Hand Surgery* 10A, 906–907 (1985)
26. Baddeley, A.D., Hitch, G.J.: Working memory. In: *Psychology of Learning and Motivation*, pp. 47–89. Academic Press, London (2000)
27. Aiken, L.R.: *Assessment of Intellectual Functioning. Perspectives on Individual Differences Series*. Springer, Heidelberg (2004)
28. Wechsler, D.: *Wechsler Adult Intelligence Scale - Revised*. Psychological Corporation, San Antonio (1981)
29. Xydias, N.: *Tests pour l'orientation et la selection professionnelle des aveugles*. Editions Scientifiques et Psychologiques (1977)

Elderly User Evaluation of Mobile Touchscreen Interactions

Masatomo Kobayashi¹, Atsushi Hiyama², Takahiro Miura³,
Chieko Asakawa¹, Michitaka Hirose², and Tohru Ifukube⁴

¹ IBM Research – Tokyo, 1623-14 Shimo-tsuruma, Yamato-shi, Kanagawa 242-8502, Japan
{mstm, chie}@jpp.ibm.com

² Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{atsushi, hirose}@cyber.t.u-tokyo.ac.jp

³ Research Center for Advanced Science and Technology, The University of Tokyo,
4-6-1 Komaba, Meguro-ku, Tokyo, 153-8904, Japan

⁴ Institute of Gerontology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{miura, ifukube}@human.rcast.u-tokyo.ac.jp

Abstract. Smartphones with touchscreen-based interfaces are increasingly used by non-technical groups including the elderly. However, application developers have little understanding of how senior users interact with their products and of how to design senior-friendly interfaces. As an initial study to assess standard mobile touchscreen interfaces for the elderly, we conducted performance measurements and observational evaluations of 20 elderly participants. The tasks included performing basic gestures such as taps, drags, and pinching motions and using basic interactive components such as software keyboards and photo viewers. We found that mobile touchscreens were generally easy for the elderly to use and a week's experience generally improved their proficiency. However, careful observations identified several typical problems that should be addressed in future interfaces. We discuss the implications of our experiments, seeking to provide informal guidelines for application developers to design better interfaces for elderly people.

Keywords: Mobile, Smartphones, Touchscreens, Gestures, Aging, Elderly, Senior Citizens, User Evaluation.

1 Introduction

Mobile phones are filling essential roles in today's societies both for the elderly and younger groups. The Ministry of Internal Affairs and Communications in Japan reported that approximately 75% of people aged six or older had their own mobile phones in 2009 [1]. At the same time, more than 70% of people aged 60-69 and 40% of people aged 70-79 used mobile phones.

It is known that senior citizens tend to use their mobile phones for relatively limited purposes (e.g., [2]). The limiting factors include hard-to-see displays, hard-to-

press buttons, and hard-to-learn procedures. Special mobile phones for the elderly (such as the *Raku-Raku Phone* from NTT DoCoMo [3]) have been developed and are widely used, and they provide such features as simplified interfaces with larger buttons. However, their limited functionality may increase the digital divide between the older and younger generations.

Touchscreen-based smartphones have the potential to address these problems. First, the finger-based intuitive interactions can be useful regardless of the age of the user. Second, touchscreen-based interfaces allow us to offer senior-friendly interfaces through software-level adaptations. Senior citizens can use the same hardware devices as younger people and also access online services and communities. This increases the social inclusion of the elderly, improving the quality of their lives and making a more sustainable society by benefiting from the power of senior citizens.

The current penetration rate of touchscreen-based smartphones for the elderly is estimated to be low. However, in the next few years they are expected to become the majority of mobile phones, since many of the new mobile phones are smartphones and the major smartphone operating systems all support touchscreens. The market of smartphone applications for senior citizens is expected to grow due to the rapidly increasing numbers of elderly mobile phone users in developed countries. Our informal preliminary survey supports this expectation, indicating that at least some senior citizens, even those who usually feel hesitation about new technologies, were interested in the latest touchscreen-based smartphones and enjoyed using gestures on their screens. Given the explosive growth of mobile phones including smartphones (even in emerging countries [4]), senior-friendly smartphone applications will be strongly desired throughout the world.

A major problem is that application developers currently have little understanding of how to design better touchscreen interfaces for elderly users since the de facto standards of basic operations on touchscreen-based smartphones, which consist of tapping, dragging, and pinching, have only been adopted in the last few years. We need frameworks and guidelines supported by empirical evidence to help develop senior-friendly interfaces. As an initial step, we need to know how the elderly interact with current touchscreen interfaces as a basis for the development of frameworks and guidelines.

We observed and measured the actions of elderly people using touchscreen smartphones. Based on previous research on senior citizens' use of traditional button-based mobile interfaces (e.g., [5][6]) and gesture-based commands on touchscreens (e.g., [7][8]), we focused on typical finger-based operations on standard devices to assess the practical use of mobile touchscreen interfaces. The goal of this study is to determine the trends and problems of mobile touchscreen interfaces that elderly users actually encounter. The tasks included: (1) controlling basic operations with gestures such as taps, drags, and pinching motions, for which we measured task completion times, analyzed their behaviors while making the motions, and asked about the users' preferences; and (2) using interactive components operated with basic gestures such as software keyboards and photo viewers, for which we simply observed their behaviors and asked for user comments. Based on the quantitative and qualitative results, we try to provide informal guidelines for application developers to design senior-friendly interfaces. Our initial experiments also suggested future research directions and new experiments.

The remainder of this paper is organized as follows. First, in Section 2 we summarize related work. Next we describe the methods of our experiments. Then we present the results of the performance measurements and review the observational analyses and the subjective feedback. We also discuss the implications of the experiments and possible future research. Finally, we summarize this work.

2 Related Work

Much research has been done to evaluate the usability for elderly people of desktop or laptop user interfaces [9][10]. Guidelines for designing accessible Web user interfaces for elderly people have been proposed [11]. In contrast, developing design guidelines for the elderly using such new interfaces as mobile terminals with touchscreen interfaces is ongoing work. Design guidelines for general touchscreen-based mobile interfaces have only recently been adopted and guidelines that consider elderly users are yet to be investigated [12].

There have been a number of user evaluation studies of touchscreen interfaces. Leonardi et al. designed a tabletop touch panel interface and found that the direct interaction metaphor was easy to understand and had a pleasing effect that attracted and motivated elderly participants in their study [13]. Lopicard et al. found two-handed touchscreen input was difficult for elderly users [14]. Stöbel et al. compared old users to young users in 42 different gesture inputs for touch surfaces and measured their speed and accuracy. They found that older users are a little slower but there was no significant difference in accuracy and suggested that older adults favor accuracy over speed [7].

Many studies of elderly users and mobile terminals have also been conducted. Seik et al. compared the performance using PDA applications between younger adults and older adults. Their results showed that both older and younger participants performed at the same level [5]. Darroch et al. examined the preferred font sizes on PDA screens, comparing older and younger adults. There were no significant differences in reading performance and accuracy between the older and younger adults, but the preferred size of the font was slightly larger for the older participants [15]. Kurniawan investigated the problems that older people face when using mobile phones and assessed some characteristics of a mobile phone for the elderly [16]. Older people are relatively passive adopters with fears of the consequences of using unfamiliar technologies, such as reduced face to face communications, or of accidents caused by careless use, such as by talking while driving. They like functions that support their declining functional capabilities.

Usability studies for new smartphones with touchscreen interfaces have just begun. Stone proposed a special free text input method for elderly people [17]. There have been some accessibility-related evaluations of mobile touchscreens for visually-impaired or motor-impaired users [18][19]. Stöbel et al. compared older users and younger users in how they interacted with touchscreens including multi-touch systems. They focused on symbolic gestures and direct manipulations. Their results showed that older users prefer direct manipulations. There were no age-related differences in direct manipulations, but considerable age differences in the use of symbolic gestures. Symbolic gestures are relatively more accepted by older users. Also, they

found no age-related differences in single-finger gestures, but younger users were more likely to use double-fingers gestures [8].

These studies give us valuable information regarding various perspectives of touchscreen interfaces for elderly people. However, few studies have focused on the up-to-date, de facto standard set of touchscreen operations, which consists of simple gestures such as tapping for selecting items, dragging for scrolling, and pinching for zooming. To derive general guidelines for designing senior-friendly interfaces on a common device, we assessed the elder's trends in performance and behaviors in the basic set of touchscreen operations.

3 Methodology

3.1 Participants

Twenty elderly Japanese in their 60s and 70s (14 females and 6 males) were paid to participate in the experiments. Their profiles are summarized in Table 1.

All of the participants had at least two years of prior experience with personal computers, except for one participant whose experience was not recorded (P15). Twelve of the participants had at least two years of prior experience with mobile phones, one was not a mobile phone user (P9), and two did not report their mobile phone experience (P1 and P15). Four of the participants had up to two years of experience with touchscreen-based mobile devices (P1, P4, P5, and P14). Given their relatively longer experience with traditional information technologies, the results of our experiments are expected to reflect the problems specific to mobile touchscreen interfaces rather than those due to the lack of general IT skills and knowledge.

Only one participant was left-handed (P18), but she interacted with the touchscreen primarily with her right hand, as did all of the right-handed participants. Some stated that they had difficulties in their daily lives due to age-related problems with their eyes (P1, P4, and P9) or age-related hearing loss (P5).

3.2 Apparatus

We used an iPad (the “large device” with a 9.7-inch, 147.8×197.1 mm multi-touch screen with 768×1024 pixel resolution, weighing 680 g) and an iPod touch (the “small device” with a 3.5-inch, 49.3×74.0 mm multi-touch screen with 640×960 pixel resolution, weighing 101 g) running iOS 4.2 in the experiments. The large device was enclosed in a case while the small device was used without a case. The experimental software described below was implemented as a native application for the small device running in the low resolution mode (320×460 pixels without a status bar at the top of the screen) at 163 ppi. On the large device, it was run in the small device emulation mode (double size) at 66 ppi. We tested the small device as a representative of mobile devices that have similar size touchscreens. We included the large device in the experiments believing that the larger touchscreen would be especially helpful for the elderly participants, since they may have visual and motor limitations that would be eased by the larger screen.

The participants sat on a chair in front of a table during the experiments. We asked them to hold the devices in a comfortable manner. As a result, the participants who

used a large device generally put it on the table while the other participants who used a small device generally held it with their left hands.

3.3 Procedure

The participants performed the series of experimental procedures twice, separated by one week, except for one participant (P11) who could not return for the second session. Each session consisted of (a) performance measurements for gesture operations and (b) realistic uses of interface components. Each session took about two hours, including a post-experiment interview. During the one-week period, each participant was asked to practice the gesture operations using the experimental software at least once per day. They were also allowed to freely use the pre-installed applications.

Performance measurements. Each participant was asked to perform four standard touchscreen operations: *tapping*, *dragging*, *pinching without panning*, and *pinching with panning*. Fig. 1 shows images of each task.

Each tapping trial started when a target, a white rounded-square button, appeared at a random location on the screen. It ended when the target was successfully tapped, meaning the finger was briefly placed on and quickly removed from the target. For each trial, the target was randomly selected from three sizes: 30, 50, and 70 pixels. The target became blue as the finger was touching the target. Once a target had been dismissed, the next target appeared until the participant had completed 30 tapping trials. Before the timed trials, the participant was allowed to have some practice trials.

We used 30, 50, and 70 pixels as the typical sizes of the keys in a software keyboard, general buttons, and icons on the home screen, respectively. We controlled the logical size in pixels, instead of the physical size, aiming to identify the problems that elder users may face in the actual use of standard mobile interfaces.

Each dragging trial started when an image (200×200 pixels) appeared at a random location on the screen. It ended when the image was moved into the target using dragging motions. The target was a green square (also 200×200 pixels) with 20-pixel-wide borders, which was always displayed at the center of the screen. The width of the borders of the target rectangle represented the tolerance, so errors up to 10 pixels in any direction were permitted for the final image location when the trial ended. The participants had some practice trials and 30 timed trials, just as with the tapping trials.

Each pinching-without-panning trial started when a variable-size square image appeared at the center of the screen. It ended when the image was adjusted to match the target by using two-finger pinching or spreading motions. The initial image was 50, 100, 300, or 400 pixels. The target was the same 200-pixel target used in the dragging trials. The final image size could range from 190 to 210 pixels. The participants again had some practice trials but only 20 timed trials.

Each pinching-with-panning trial started when a square image appeared at a random location on the screen. It ended when the image's size and location had been adjusted to match the same target rectangle. This required two-finger pinching and spreading motions and one- or two-finger dragging. The initial image sizes and the target size were the same as in the "without panning" condition. The participants performed some practice trials and 20 timed trials.

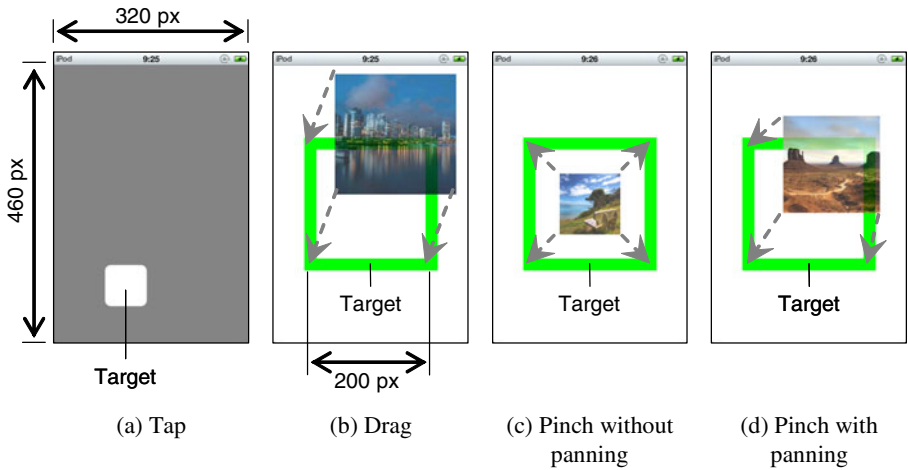


Fig. 1. Our experimental software was designed to practice and assess four standard touch-screen operations: (a) tapping, (b) dragging, (c) pinching without panning, and (d) pinching with panning. The text, lines, and arrows are annotations, not displayed in the actual experiments.

Realistic use. The participants were asked to use two standard interactive components: *photo viewer* and *software keyboard*. We chose these two components as representative of fundamental tasks in online communications: browsing and entering information. Also, they require the use of all of the tested gestures. First each participant was briefly shown how to use each component. Then they were asked to use the component for about 20 minutes with informal tasks such as finding photos that contain more than ten people and copying some printed reference sentences. We also briefly introduced them to some frequently used applications such as cameras, music players, e-book readers (for magazines and newspapers), and games.

3.4 Design

For the performance measurements, the *device* (large or small device) was a between-participant factor, where P0-P9 used the large device and P10-P19 used the small device. The *week* (1 or 2) was a within-participant factor. Other within-participant factors included the *target size* for tapping tasks and *initial size* for pinching tasks. We measured the *task completion time* as the primary dependent variable. We also investigated the participants' behaviors in depth and asked about their preferences.

For the realistic use experiment, we observed the participants' behaviors and then asked them for comments.

3.5 Hypotheses

Here are the main hypotheses we sought to test in this study:

1. Basic touchscreen operations are easy for the elderly to perform without training.
2. A week's practice will improve the performance.

3. The simplest gesture, tapping, is the easiest to perform.
4. Pinching with panning is the most difficult to perform because of its complexity.
5. The larger screen is preferred and more efficient for tapping, because the targets are larger, but the smaller screen is better for dragging and pinching operations, requiring smaller finger movements.
6. For the interactive components, the participants face problems similar to those of traditional user interfaces, such as unclear instructions and unclear indicators of the current state.

Table 1. Twenty senior citizens participated in the experiments. P0-P9 tested a large screen device while P10-P19 tested a small screen device

#	Age	Gender	Dominant hand	Prior experiences			Difficulties in vision/hearing
				PC	Mobile	Touch-screen	
P0	70s	Female	Right	10 years	6 years	none	-
P1	70s	Male	Right	10 years	-	1 year	Vision
P2	70s	Female	Right	10 years	6 years	none	-
P3	70s	Male	Right	2 years	5 years	none	-
P4	70s	Female	Right	10 years	5 years	once	Vision
P5	70s	Female	Right	8 years	2 years	2 years	Hearing
P6	70s	Male	Right	15 years	3 years	none	-
P7	70s	Male	Right	23 years	15 years	none	-
P8	70s	Female	Right	6 years	3 years	none	-
P9	60s	Female	Right	6 years	none	none	Vision
P10	60s	Female	Right	7 years	8 years	none	-
P11	60s	Female	Right	10 years	10 years	none	-
P12	70s	Female	Right	8 years	3 years	none	-
P13	70s	Male	Right	10 years	5 years	none	-
P14	70s	Male	Right	15 years	8 years	once	-
P15	60s	Female	Right	-	-	none	-
P16	60s	Female	Right	6 years	6 years	none	-
P17	60s	Female	Right	15 years	10 years	none	-
P18	70s	Female	Left	10 years	11 years	none	-
P19	60s	Female	Right	10 years	5 years	none	-

4 Basic Gesture Performances

We found no significant effects on the task completion times due to the age, gender, dominant hand, prior experience, or difficulties in vision/hearing. This does not prove that these factors cannot affect the performance because we did not control for them in this experiment. Their precise effects or irrelevance should be assessed in future research.

4.1 Taps

Fig. 2 shows the average task completion times for tapping operations for each target size. For the first week, the average times were 0.77 seconds and 1.02 seconds for

large and small devices, respectively. For the second week, they were 0.72 seconds and 0.94 seconds. A week's practice reduced the task completion times by 7% and 9% for large and small devices, respectively.

Analysis of variance showed significant main effects on the task completion times due to the device ($F_{1,1158} = 14.43, p < .001$) and the target size ($F_{2,1158} = 28.81, p < .001$). It also showed a significant interaction effect between the device and the target size ($F_{2,1158} = 8.96, p < .001$). There were no significant main effects of the week's practice ($F_{1,1158} = 1.12, p > .05$). A *post hoc* analysis found that tapping 30-pixel targets on the small device was significantly more difficult than the other conditions. It took about twice as long to tap the 30-pixel target compared to the 50- or 70-pixel targets.

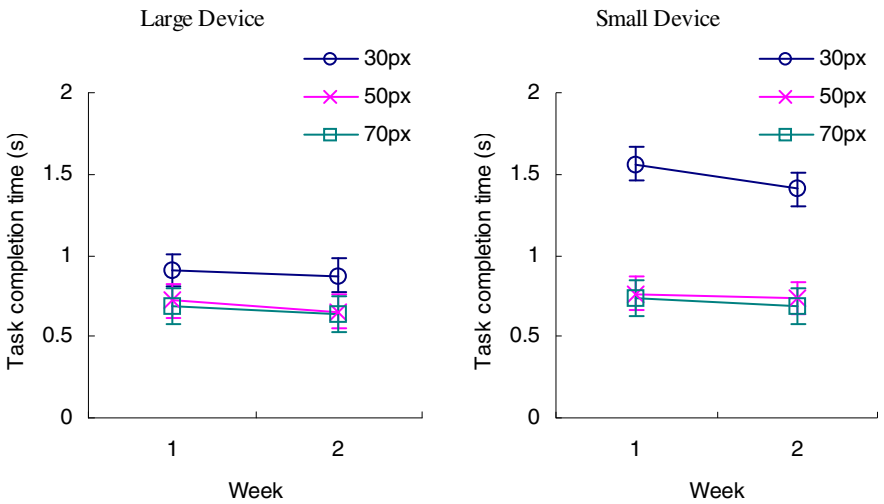


Fig. 2. Tapping task completion times for each target size with standard errors before and after a week's practice

4.2 Drag

Fig. 3 shows the average task completion times for dragging operations. For the first week, the average times were 2.09 seconds and 2.17 seconds for the large and small devices, respectively. For the second week, they were 1.59 seconds and 1.77 seconds. A week's practice reduced the task completion times by 24% and 18% for the large and small devices, respectively.

Analysis of variance showed significant main effects from the week's practice ($F_{1,1138} = 50.83, p < .001$) and for the device ($F_{1,1138} = 4.52, p < .05$) on the task completion times. There were no significant interaction effects.

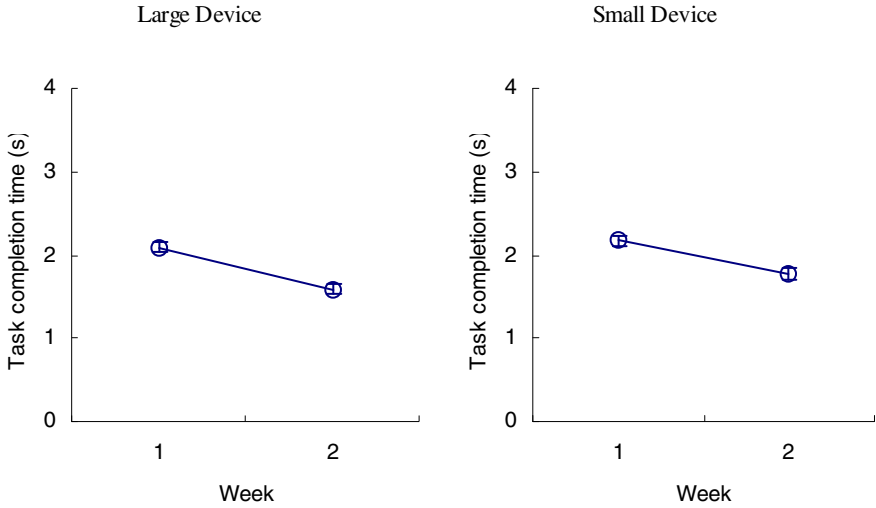


Fig. 3. Task completion times for dragging with standard errors before and after a week's practice

4.3 Pinching and Spreading without Panning

Fig. 4 shows the average task completion times for zooming in and out operations for each initial image size. In the first week, the average times were 2.69 seconds and 2.75 seconds for the large and small devices, respectively. For the second week, they were 1.92 seconds and 1.99 seconds. A week's practice reduced the task completion times by 29% and 28% for the large and small devices, respectively.

Analysis of variance showed significant main effects on the task completion time of the week's practice ($F_{1,753} = 100.15, p < .001$) and from the initial size ($F_{3,753} = 20.87, p < .001$). It also showed a significant interaction effect between the device and the initial size ($F_{3,753} = 6.51, p < .001$). There were no significant main effects of the device ($F_{1,753} = 0.63, p > .05$). A *post hoc* analysis found that spreading motions were more difficult than pinching motions, especially on the small device. It took 3.13 seconds to zoom in to resize the image from 50 pixels to 200 pixels while it took only 2.19 seconds to zoom out from 400 pixels to 200 pixels, even though the required amount of finger movement on the screen was larger for zooming out. Some participants commented that spreading (zooming-in) motions were more difficult than pinching (zooming-out) motions, reinforcing these results.

4.4 Pinching and Spreading with Panning

Fig. 5 shows the average task completion times for zooming with panning operations for each initial image size. In the first week, the average times were 4.57 seconds and 5.01 seconds for the large and small devices, respectively. In the second week, they were 3.20 seconds and 3.60 seconds. A week's practice reduced the task completion times by 30% and 29% for large and small devices, respectively. The times were

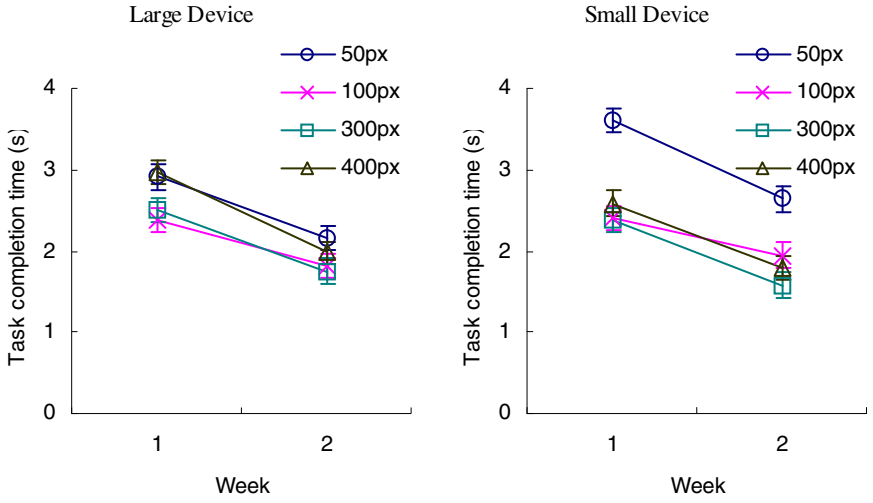


Fig. 4. Task completion times of pinching without panning with standard errors before and after a week's practice

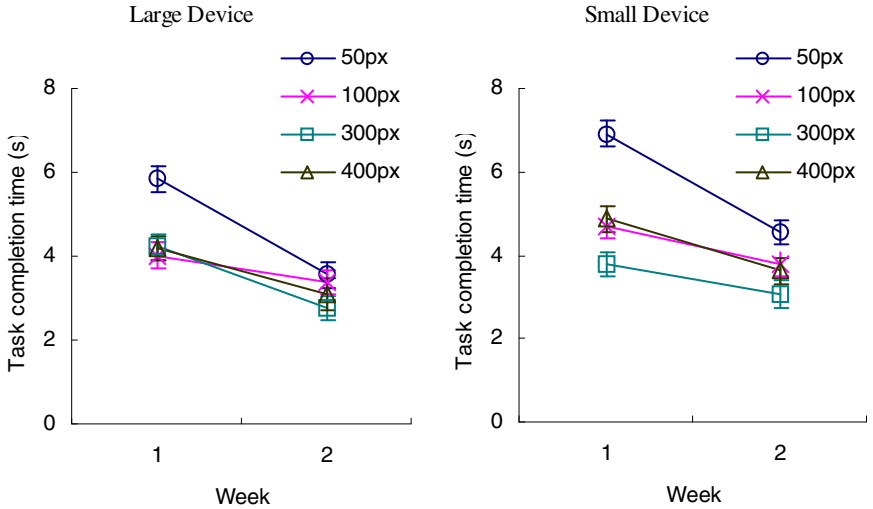


Fig. 5. Task completion times for pinching with panning with standard errors before and after a week's practice

approximately the total of the times for separately dragging and pinching without panning. This seems to confirm that the participants often used a 2-phase approach, where they first dragged the image to the center of the screen and then zoomed in or out to fit the target rectangle.

Analysis of variance showed significant main effects on the task completion times of the week ($F_{1,738} = 76.71, p < .001$), the device ($F_{1,738} = 11.92, p < .005$), and the initial size ($F_{3,738} = 18.85, p < .001$). It also showed a significant interaction effect between the week and the initial size ($F_{3,738} = 5.09, p < .005$). A *post hoc* analysis found that a week's practice greatly improved the performance, especially for zooming in from 50 pixels to 200 pixels (6.38 seconds and 4.06 seconds for weeks 1 and 2, respectively).

5 Observational Findings

5.1 Why Is Tapping a Small Target So Difficult?

To help explain why tapping 30-pixel targets showed much worse performance on the small device, Fig. 6 shows the locations where the participants actually touched in relative coordinates with respect to the target location. The average x-locations were off center by +12.9 ($SD = 10.0$) and +8.28 ($SD = 24.2$) pixels for Weeks 1 and 2, respectively. In contrast, the corresponding values for the y-axis were only +2.0 ($SD = 9.7$) and +1.8 ($SD = 31.2$) pixels. The participants clearly tended to touch the right side of the target. The size and location of the target did not affect this tendency. The gap between intended and actual touch locations could be explained by the fact that all of the participants used their right hands to perform the operations. The gap in the (x, y) range was similar for the 50- and 70-pixel targets on the small device while it was smaller for the large device. The touch locations mostly ranged from (-25, -25) to (+25, +25). As a result, the 30-pixel targets frequently caused errors and increased the task completion times, while the larger targets had few errors. For 30-pixel targets on the small device, the error rate was 39%, where an error means the participant tapped the screen outside of the target at least once during the trial. This error rate was much worse than the acceptable 4% value used in most human-computer interaction (HCI) literature. The error rates were 6.5% and 5.6% for 50- and 70-pixel targets, respectively, on the small device. The respective values were 13.6%, 1.4%, and 1.7% for 30-, 50-, and 70-pixel targets on the large device.

The distribution of touch locations also indicates why the performance was little improved after a week's practice. As shown in Fig. 6, the variance of the locations in Week 2 was actually larger than in Week 1. Our observations and the participants' comments showed that this was caused by a change in the participants' error correction strategy. Due to the gap between the intended and actual touch locations, the participants often tapped outside the target and saw that nothing happened. In that case, in Week 1 they tended to tap the same place until they received the expected feedback. In Week 2, they seemed to change their strategy and tapped at different places, now being aware of the gap between the intended and actual touch locations and trying to correct for it. However, this strategy often resulted in an error by overshooting the target, especially for small 30-pixel targets on the small device (Fig. 7-a). In addition, we found another type of strategy change. In Week 2, some participants tried to vary the speed and pressure by slowing down or speeding up their finger movements if they failed to tap the target on the first attempt. The average contact times with the screen when tapping were 105 ms ($SD = 43$ ms) and 114 ms

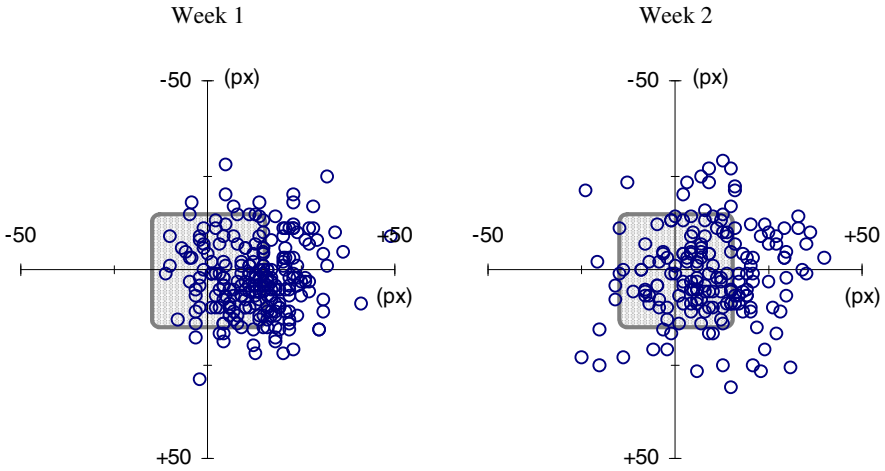


Fig. 6. The distribution of touch locations relative to the location of the 30-pixel target on the small device before and after a week’s practice. The target area is shown as a rounded square. The participants tended to tap the right side of the target. The variation of touch locations was larger after the week’s practice.

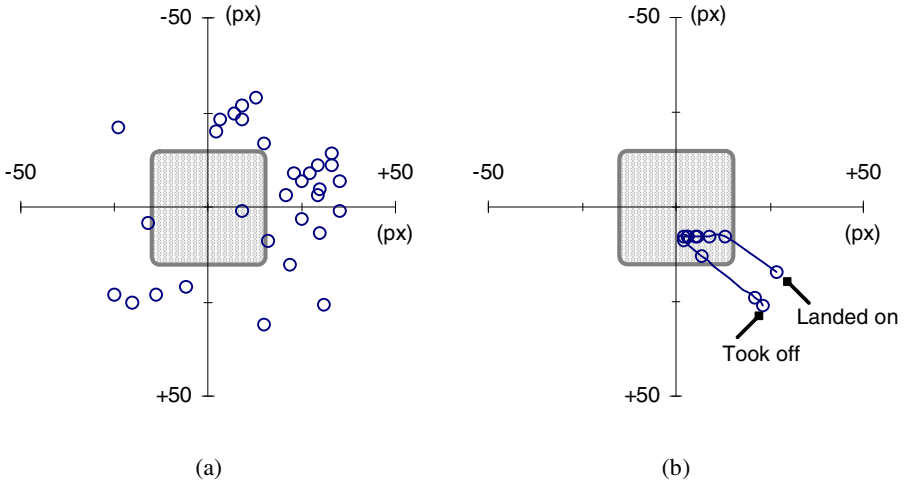


Fig. 7. (a) Due to repetitive overshooting, a participant touched the screen 32 times to tap a small target (P12, Week 2). (b) Adding extra pressure (i.e., “pressing”) on the screen while tapping can cause an irregular trajectory (P13, Week 2).

($SD = 76$ ms) for Weeks 1 and 2, respectively. This increased-pressure strategy often caused other problems. First, an overly slow tap can be recognized as a “hold” operation by many applications. Second, exerting more pressure on the screen during a slow motion can cause the finger to draw a random path, which is then recognized as

a drag (Fig. 7-b). The low visibility of the visual feedback seemed to make these problems even worse, since a dynamic change of the target color was often obscured by the finger when tapping a small (30-pixel) target. Due to these factors, five of the participants (P7 who used the large device and P12, P13, P15, and P16 who used the small device) actually became slower after a week's practice.

5.2 Problems in Applications

The problems with tapping small targets were frequently observed during the use of the practical interactive components. For example, the participants failed to tap the proper key on the software keyboard or failed in moving the cursor to a specific position in a text input area. In addition, we found that the participants faced general user interface problems, which seem to be common with traditional user interfaces.

First, the participants were often confused due to unclear instructions. A typical gesture-based interface on a touchscreen lacks textual labels and menus to show "what is possible with this application". It also provides few explicit instructions for each gesture. Even though an interface is carefully designed so that it is natural and intuitive for users to control many of its functions, the participants tended to stick to the few functions that the experimenters had explained. They were sometimes confused when they invoked an unexpected function with an unintentional gesture. A participant commented that he wanted standard menus as with a personal computer (PC). Though the participants were experienced PC users and familiar with standard QWERTY keyboards, they often could not find the backspace and shift keys on the software keyboard because those keys did not have the textual labels "Backspace" and "Shift".

The participants were also confused when an application has multiple modes without an indication of the current mode. For example, the photo viewer provides multiple thumbnail views such as by folder, by event, and by person, but it was difficult to tell which view was active. The software keyboard was more problematic. Since Japanese needs to input thousands of Chinese-related characters as well as characters using three alphabetic sets, Arabic digits, and symbols, the interface requires frequent mode-switching between predictive input [20], 5-touch input [21], and a QWERTY keyboard. The participants frequently lost track of their current mode, especially when another mode had a similar look-and-feel. One participant explicitly complained that there was no textual indication of the current mode.

6 Subjective Feedback

In contrast to our expectations, the participants commented that dragging and pinching were preferred and easier to do than tapping. In particular, most of the participants stated that dragging was the easiest operation to perform. They found that pinching was also easy, though some reported that pinching with panning was somewhat difficult.

As regards the device, the tradeoff between the size and weight was mentioned. Many participants said that the large device was too heavy to carry while the small device was too small to read. Some female participants wanted a middle-sized, 7-inch device that would fit into their handbags.

When it came to applications, the participants wanted to use various types of applications, such as cameras and photo viewers, newspapers, e-books, and games. Many of them stated that they especially wanted smartphones when they had a few extra minutes, such as when riding the train.

The participants frequently commented that using touchscreen interfaces was “enjoyable”. For example:

It is fun to flip through photos [in the photo viewer]. (P5)

I want to do nothing but use this [a touchscreen]. (P14)

These comments confirm the findings of previous research [22]. The enjoyability is one of the most important factors in encouraging the elderly to continue participating in society via information technologies. We need more investigation of the factors influencing why they enjoyed using these devices. Perhaps it is part of the nature of touchscreen-based interfaces or the participants were excited by using new devices.

7 Discussion

The experiments partly confirmed and partly rejected our hypotheses. The first hypothesis was generally confirmed. The participants, even those who had never used touchscreens, performed the gesture-based operations reasonably well, except for tapping on small targets. The second hypothesis was partly confirmed. A week’s practice significantly improved the performance for dragging and pinching operations, though there were no significant effects for tapping. The third and fourth hypotheses were rejected. The participants stated that dragging and pinching were easier and more comfortable than tapping. Though some of them found that pinching with panning was a little difficult, their performance was not seriously hindered by the difficulty. The fifth hypothesis was partly confirmed. It was surprising that even for dragging and pinching, as well as for tapping, the larger screen outperformed the smaller screen, even though it required more than twice the amount of finger movement on the screen. However it was reported that the device with the large screen was too heavy to carry and use outdoors. The last hypothesis was also confirmed.

7.1 Design Implications

Based on the experimental findings, we suggest these informal guidelines for application developers seeking to design better interfaces for the elderly.

Use larger targets (8 mm or larger in size). The elderly tend to make errors when tapping a small target, such as a 30-pixel button on a small screen. In the experiment, the touch locations were mostly distributed within 8 mm on the physical screen regardless of the device and the target size. Thus, interactive objects such as buttons, icons, and clickable text should at a minimum be larger than 8 mm. Note that some common components in the latest smartphone interfaces, such as keys in a software keyboard and a “back” button in the upper-left corner of a small device, conflict with this guideline. With targets located close to each other without spacing (e.g., software

keyboard), the size should be much larger to avoid invoking adjacent targets. This guideline confirms the result of previous work [23], which found that the speed and error rates were much worse when tapping buttons smaller than approximately 8 mm.

Address the gap between intended and actual touch locations. The elderly tend to miss their intended targets due to parallax and the large contact area of each finger. Providing a calibration mechanism might be a solution. However, practical calibration is difficult because of the nature of handheld devices. Users frequently tilt and rotate their devices depending on their situations. Another solution might include providing appropriate visual feedback to indicate where the users touched the screen, even when they have missed all of the interactive targets.

Consider using drag and pinch gestures rather than taps. The elderly tend to prefer dragging and pinching operations over tapping. This is a new finding that was not observed in an earlier experiment on multi-touch interactions for elder users [8]. It is also the opposite of the result for motor-impaired users [18]. Though applications need not avoid using drag and pinch operations, the application should provide instructions and clear visual clues to show which gesture invokes which function. Since typical functions invoked by dragging and pinching provide no visual clues, such as flipping and zooming photos, most participants in our experiment could not find those functions without guidance.

Explicitly display the current mode. Some elderly users are less likely to notice changes of the modes and can become confused. Applications should avoid multi-mode interfaces as much as possible. If a multi-mode design is needed, then there should be explicit feedback about mode-changes and a display of the current mode. The indicators of mode changes should be persistent and large, since the elderly users may fail to notice short alerts or small changes in the look-and-feel. Also, the feedback should be readable as they tend to have difficulty in interpreting the meanings of symbolic representations.

7.2 Future Work

A limitation of this study is that we did not have younger participants as a reference group. We focused on investigating how the elderly actually interact with mobile touchscreens and how practice improves their proficiency, as a first step towards guidelines and frameworks to develop senior-friendly interfaces. Hence in-depth analyses of aging effects were beyond our scope. An informal experiment we conducted with two young participants (2 females, 24 and 31 years old), in which they performed touchscreen operations 3%-49% faster than the senior participants, indicated that younger users interact more rapidly with touchscreen interfaces. However, the general trends were similar. For example, considering the result of previous research [23], the minimal target size that a user can tap with good speed and accuracy seems to be around 8 mm for both younger and older people. Additional investigation of the similarity and differences between younger and older users in their performances, behaviors, and preferences would help in designing more universal interfaces that satisfy users of all ages.

In addition to the comparison with younger users' performance, more investigation is needed to understand the trends and problems we encountered. Are they caused by physical and cognitive disabilities such as attention and memory loss related to aging? How do the experiences with other technologies affect performance? A study with more control of the variables such as age, disability, and experience with technologies may discover clues towards more practical solutions for specific user groups.

8 Conclusion

We conducted performance measurements and observational evaluations to assess standard mobile touchscreen interfaces when used by the elderly. Our participants were 20 Japanese in their 60s and 70s. The tasks included (1) controlling basic operations with gestures such as taps, drags, and pinching motions and (2) using basic interactive components such as software keyboards and photo viewers. The results show that touchscreen mobile interfaces are preferred and not too difficult to use, even by the elderly. A week's practice significantly improved the performance in dragging and pinching, but did not significantly affect tapping. We identified several typical problems, such as mismatches between the user's visual target and the touched position as detected by the sensor. We discussed the implications of the experiments on the design of better interfaces for the elderly and considered future research directions. This and future studies will provide the basis for the development of senior-friendly user interfaces.

Acknowledgments. This research was partially supported by the Japan Science and Technology Agency (JST) under the Strategic Promotion of Innovative Research and Development Program. We thank Ms. Yoriko Ohbayashi, Mr. Sternly K. Simon, Mr. Daisuke Ochiai, Mr. Daisuke Sato, and Dr. Susumu Harada for their assistance. We also thank all of the participants in our experiments.

References

1. Ministry of Internal Affairs and Communications in Japan: Communications Usage Trend Survey (2010)
2. Office for National Statistics in the U.K.: Use of ICT at Home (2007)
3. Universal Design Initiatives, http://www.nttdocomo.com/about/csr/universal_design/
4. ICT Statistics, <http://www.itu.int/ITU-D/ict/statistics/ict/>
5. Siek, K.A., Rogers, Y., Connelly, K.H.: Fat Finger Worries: How Older and Younger Users Physically Interact with PDAs. In: Costabile, M.F., Paternó, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 267–280. Springer, Heidelberg (2005)
6. Ueda, K., Okochi, N., Chikawa, A., Ito, A., Hiramatsu, Y., Ifukube, T.: Usability Evaluation of Mobile Phone for the Elderly: The Effect of Properties of the Button. In: Human Interface 2009, pp. 673–676 (2009) (in Japanese)
7. Stöbel, C., Wandke, H., Blessing, L.: Gestural Interfaces for Elderly Users: Help or Hindrance? In: Kopp, S., Wachsmuth, I. (eds.) GW 2009. LNCS, vol. 5934, pp. 269–280. Springer, Heidelberg (2010)

8. Stöbel, C., Blessing, L.: Mobile Device Interaction Gestures for Older Users. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction (NordiCHI 2010): Extending Boundaries, pp. 793–796 (2010)
9. Arias-Torres, D.: The Design and Evaluation of a Pen-Based Computer Interface for Novice Older Users. In: Proceedings of the 7th Mexican International Conference on Computer Science (ENC 2006), pp. 142–150 (2006)
10. Mahmud, M., Kurniawan, H.: Involving Psychometric Tests for Input Device Evaluation with Older People. In: Proceedings of the 19th Conference of the Computer-Human Interaction Special Interest Group (CHISIG) of Australia on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future (November 21–25, 2005)
11. Web Accessibility and Older People: Meeting the Needs of Ageing Web Users, <http://www.w3.org/WAI/older-users/>
12. iOS Human Interface Guidelines, <http://developer.apple.com/library/ios/documentation/userexperience/conceptual/mobilehig/>
13. Leonardi, C., Albertini, A., Pianesi, F., Zancanaro, M.: An Exploratory Study of a Touch-Based Gestural Interface for Elderly. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction (NordiCHI 2010): Extending Boundaries, pp. 845–850 (2010)
14. Lepicard, G., Vigouroux, N.: Touch Screen User Interfaces for Older Subjects: Effect of the Targets Number and the Two Hands Use. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010. LNCS, vol. 6180, pp. 592–599. Springer, Heidelberg (2010)
15. Darroch, I., Goodman, J., Brewster, S., Gray, P.: The Effect of Age and Font Size on Reading Text on Handheld Computers. In: Costabile, M.F., Paternó, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 253–266. Springer, Heidelberg (2005)
16. Kurniawan, S.: Older People and Mobile Phones: A Multi-Method Investigation. International Journal of Human-Computer Studies 66(12), 889–901 (2008)
17. Stone, R.G.: Mobile Touch Interfaces for the Elderly. In: Proceedings of ICT, Society and Human Beings (July 22–24, 2008)
18. Guerreiro, T., Nicolau, H., Jorge, J., Goncalves, D.: Towards Accessible Touch Interfaces. In: Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2010), pp. 19–26 (2010)
19. McGookin, D., Brewster S., Jiang, W.: Investigating Touchscreen Accessibility for People with Visual Impairments. In: Proceedings of the 5th Nordic Conference on Human-Computer Interaction (NordiCHI 2008): Building Bridges, pp. 298–307 (2008)
20. Masui, T.: POBox: An efficient text input method for handheld and ubiquitous computers. In: Gellersen, H.-W. (ed.) HUC 1999. LNCS, vol. 1707, pp. 289–300. Springer, Heidelberg (1999)
21. Macdonald, A.S.: The UD Phenomenon in Japan: Product Innovation through Universal Design. In: Stephanidis, C. (ed.) HCI 2007. LNCS, vol. 4554, pp. 224–233. Springer, Heidelberg (2007)
22. Chikawa, A., Ueda, K., Okochi, N., Ito, A., Hiramatsu, Y., Ifukube, T.: Usability Evaluation of Touch Screen Mobile Phone for the Elderly. In: Human Interface 2009, pp. 677–680 (2009) (in Japanese)
23. Lee, S., Zhai, S.: The Performance of Touch Screen Soft Buttons. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI 2009), pp. 309–318 (2009)

BrailleType: Unleashing Braille over Touch Screen Mobile Phones

João Oliveira, Tiago Guerreiro, Hugo Nicolau,
Joaquim Jorge, and Daniel Gonçalves

INESC-ID / Technical University of Lisbon, Portugal
jmgdo@ist.utl.pt, {tjvg,hman,jaj,djvg}@immi.inesc-id.pt

Abstract. The emergence of touch screen devices poses a new set of challenges regarding text-entry. These are more obvious when considering blind people, as touch screens lack the tactile feedback they are used to when interacting with devices. The available solutions to enable non-visual text-entry resort to a wide set of targets, complex interaction techniques or unfamiliar layouts. We propose BrailleType, a text-entry method based on the Braille alphabet. BrailleType avoids multi-touch gestures in favor of a more simple single-finger interaction, featuring few and large targets. We performed a user study with fifteen blind subjects, to assess this method's performance against Apple's VoiceOver approach. BrailleType although slower, was significantly easier and less error prone. Results suggest that the target users would have a smoother adaptation to BrailleType than to other more complex methods.

Keywords: Blind, Braille, Mobile Devices, Text-Entry, Touch screens.

1 Introduction

Long are the days where mobile phones were considered a luxury, as they are now an integral part of our daily lives. Progressively more powerful, these devices allow an ever growing set of functionalities, meant not only for communication purposes, but also for productivity and leisure. Even though they are becoming increasingly ubiquitous, they are still far from being accessible to everyone. Disabled target groups, such as blind people, still struggle with these visually demanding devices.

The emergence and success of touch screen devices, which are gradually replacing the traditional keypad ones, can pose a daunting future to the blind, as several challenges arise. Interaction with touch screens is even more demanding from a visual standpoint, as familiar and more easily identifiable input methods, such as keyboards, are replaced by their virtual onscreen counterparts. The lack of tactile feedback and the physical stability offered by keypads can turn selection of targets much harder, or even render the blind user virtually lost. Besides the fact that these devices are usually devoid of physical buttons, the interface is constantly changing from screen to screen, depending of context, making it hard to navigate and access the desired content.

The aforementioned problems make touch screen text-entry a major challenge for a blind person. Several single-touch, as well as multi-touch solutions, based on hitting

targets, gestures or a combination of both have been proposed to address this problem. Yfantidis and Evreinov [4] developed an interface based on simple unidirectional single-finger gestures, regardless of position, to select characters. Although this method eliminates the need to search for targets, it is very dependent on the gesture-finger orientation, which can become troublesome since there is not a standard way to hold devices and maintain orientation as it was envisioned. Apple's VoiceOver¹ text-entry solution, relies on a soft keyboard in which the users focus the desired key by touching it, and enters it by split-tapping or double tapping anywhere. On the strong side, it enables the blind user to input text similarly to a sighted person with a simple screen reading approach. On the other hand, VoiceOver typically uses a QWERTY soft keyboard layout, hence it features a large number of targets, making it difficult to find a specific letter especially if the user is not familiar with computer keyboards. In a similar fashion, Bonner et al. system No-Look Notes [1], also uses targets and selection through multi-touch techniques like split-tapping. However No-Look Notes minimizes the number of targets on screen by dividing it into eight sectors containing groups of characters, just like the standard 12-key phone keypad. Split-tapping one of these segments brings another screen, with the corresponding letters arranged in new segments, for the user to select the same way he did previously. This solution improves on some of VoiceOver's problems, but still presents users with an unfamiliar and inconsistent layout, changing from a circular pie menu layout in the character group screen, to a vertical list in the character selection screen. Furthermore, the use of split-tapping among other gestures, as simple as they are, can be quite difficult to people with dexterity and coordination problems.

In light of these problems, we present BrailleType, a single-touch text-entry system for touch screen devices. BrailleType allows the blind user to enter text as if he was writing Braille using the traditional 6-dot matrix code. The Braille system is simple yet powerful, as any character, including accentuated letters, can be made through the combination of six or less dots. BrailleType takes advantage of this knowledge to allow the user to input text, resorting to a single screen composed of 6 targets representing the Braille matrix. This paper details BrailleType as well as its evaluation through a user study with fifteen blind subjects.

2 BrailleType

The Braille system was devised by Louis Braille in 1825 as a method of writing and reading for blind people. Although its use is declining due to the use of electronic text in conjunction with assistive software such as screen readers, it is still a widely used method and paramount to the daily lives of blind people.

Each Braille character or cell is represented by combinations of dots on a 3x2 matrix (Figure 1). There are 63 possible combinations, making it possible to represent alphabet letters, accentuated letters, punctuation, numbers, mathematical symbols or even musical notes.

The use of the Braille alphabet, common knowledge for many blind users, to enter text on a mobile keypad device was already explored by Guerreiro et al. on

¹ <http://www.apple.com/accessibility/iphone/vision.html> (Last visited on 07/04/2011).

BrailleTap [2]. In that system, the Braille cell was mapped to different keypad buttons (keys ‘2’, ‘3’, ‘5’, ‘6’, ‘8’ and ‘9’ on the mobile phone). Each press on those keys would fill the matrix with the corresponding dots, accepting the character through the press of another button (key ‘4’).

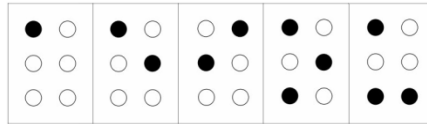


Fig. 1. From left to right: the vowels: ‘a’, ‘e’, ‘i’, ‘o’, ‘u’ in Braille

BrailleType comes as an adaptation of this approach to the now widely popular touch screen devices. Herein, the touch screen serves as a representation of the Braille cell, having six large targets representing each of the dots positions. These targets were made large and mapped to the corners and edges of the screen to allow for an easy search (Figure 2). Since they are also arranged accordingly to the known and expected Braille cell, the targets become spatially easy to find.

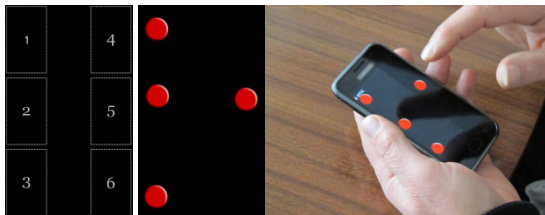


Fig. 2. BrailleType main screen on the left (visual representation of the six target zones was added for illustration). Middle screen shows the letter ‘r’ marked and ready to be accepted. The image on the right shows a user writing the letter ‘r’ with BrailleType.

In order to keep this system simple and undemanding, multi-touch techniques were avoided. All interactions with BrailleType are made through single finger input. Whenever a user presses or drags a finger to a new target, the corresponding dot number is audibly announced, but the target is not immediately selected. A small timer was implemented to prevent involuntary selections from the user. This timer can be easily configured, allowing for a more experienced user to decrease or eliminate the waiting time.

To mark a dot in the Braille cell, the user must touch the desired target, and wait for an audio confirmation cue. Repeating the process on a dot already marked, would remove it and inform the user through audio feedback. After marking all the necessary dots for a Braille character, in whichever order the user desires, a double-tap in any part of the screen accepts it. A space is made by entering an empty Braille cell. If the user tries to accept an incorrect combination of dots, the Braille matrix is cleared and an error sound is played.

A swipe to the left clears the Braille cell if one or more dots were already marked or deletes the last entered character if the matrix was empty.

This method seeks to provide a less stressful first approach with touch screen devices by reducing the number of onscreen targets. By reducing the number of errors and enabling the user to succeed, we augment their confidence to go farther. Also, we intend to take advantage of the capabilities of those who use Braille on a regular basis, but also enable those who do not to learn or maintain Braille usage through simple daily interactions (e.g., writing text messages).

3 Evaluation

We conducted a comparative evaluation of BrailleType and VoiceOver with the target population. On two sessions, one for each system, participants first learned how to use both text-entry methods and had the opportunity to interact and practice with them. After this tutorial, participants were asked to write a set of sentences and end the session with a brief questionnaire. Any input made by the users during the evaluation was logged for later analysis. We focused both on the efficiency and effectiveness of the participant's writing, through the analysis of WPM and MSD error rates metrics.

3.1 Participants

Fifteen blind people (light perception at most) were recruited from a formation centre for the visually impaired. The evaluations took place in a quiet office room of this center. The participant group was composed of 10 males and 5 females, with ages ranging from 24 to 68 years old (averaging 45).

With more or less difficulty, the entire target group, with the exception of two subjects, writes text messages on their mobile phones with the help of screen readers. These two participants did not have mobile phones with screen readers and stated that they had never written a text message, restricting the phone use to placing and receiving calls. In terms of touch screen devices, only one user had made contact with this technology previously.

All of the participants knew the Braille alphabet and had had previous contact with a computer keyboard, thus being familiar to some extent with the QWERTY layout.

3.2 Material

For this experimental task we used the Samsung Galaxy S touch screen device, which runs on Google's operating system Android. This device has a 4 inch capacitive touch screen with multi-touch support, and although the screen does not occupy the entire device, since it has upper and bottom non-touch sensitive zones, no tactile upper and bottom boundaries were created.

BrailleType and a text-entry method identical to VoiceOver, in both keyboard layout and interaction methods, were implemented as Android applications. All audio feedback was given using SVOX Classic TTS. An application to manage both text-entry methods, user sessions and sentences required to type was also implemented. This application informed which sentence to type and logged all the participants' interactions (focus and entry), so it could be analyzed afterwards.

3.3 Procedure

This user study was composed of two sessions per user, each one focusing on one of the two text-entry methods, BrailleType and VoiceOver. The order in which the sessions were undertaken was decided randomly to counteract order effects.

In both sessions, with the help of the experimenter, participants started by learning the system and interacting with it for a minimum of 10 minutes and a maximum of 15 minutes. Each possible action was exemplified and taught to the participants and, as they trained, they were encouraged to ask questions and allay all doubts. If by the end of 15 minutes the participant was unable to write his name or a simple, common four-letter word, the evaluation was not continued. For the BrailleType method, two different timeout values for target selection were used. An introductory one of 1250 ms after entering the target zone, and a second one of 800ms. This last value was set when the experimenter felt the participant was confident enough and understood the basic actions of the method. All participants performed the test with the second value.

After the practice phase, participants were instructed to write a set of sentences as fast and accurately as they could, without the need to put any accentuation or punctuation. Each trial consisted of 5 phrases, each with 5 words with an average size of 4.48 characters. These sentences were extracted from a written language *corpus*, and each one had a minimum correlation with the language of 0.97. The sentences' selection was managed by the application and randomly presented to the user to avoid order effects.

All focused and entered characters were registered by the application during the evaluation. Since we wanted to focus on the number of errors participants made, the option to delete a character was locked. If a participant made a mistake or was unable to input a certain letter, she/he was told not to worry and simply carry on with the next character. It was made clear to all participants that we were testing the system and not their writing skills. When participants finished each sentence, the device was handed to the experimenter to load the next random sentence and continue with the evaluation. All sentences were read aloud by the experimenter and then repeated by all participants to ensure that they understood them correctly.

After writing all 5 test phrases, the session ended with a brief questionnaire to collect the participants' opinions on the text-entry method. All these steps were repeated for both techniques.

4 Results

Our goal in this study was to assess, in an exploratory evaluation, how BrailleType stands against a VoiceOver alike solution, in regards to speed, accuracy and preference.

4.1 Text-Entry Speed

Concerning speed, we used the WPM text entry measure, calculated as $(\text{transcribed text} - 1) * (60 \text{ seconds} / \text{time in seconds}) / (5 \text{ characters per word})$ [3]. Time to input

each sentence was measured from the moment the first character was entered to the last. Figure 3 shows the performance of each participant, speed wise, on both methods. Some participants had similar WPM on both methods, but VoiceOver was generally faster, with an average of 2.11 WPM against BrailleType's 1.45 WPM.

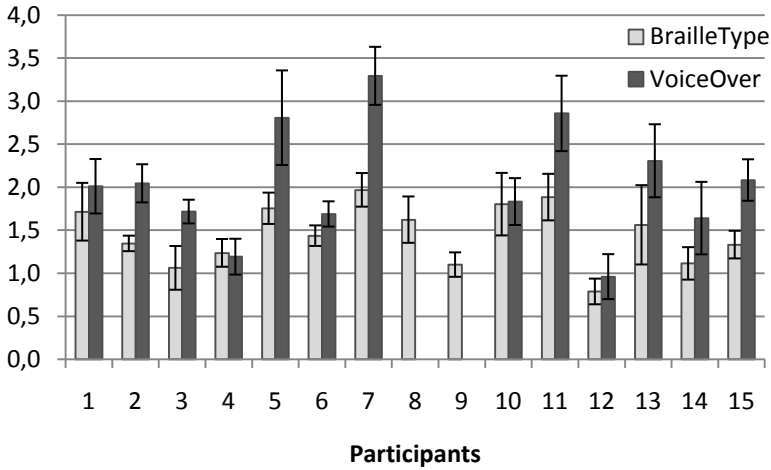


Fig. 3. WPM (average) for each participant on each system

Participants 8 and 9 after the 15 minutes of the practice session were still struggling with the VoiceOver alike method, so they did not perform the test. Given the normality of the data (according to the Shapiro-Wilk test), a One-way ANOVA was applied to confirm that VoiceOver-alike was significantly faster than BrailleType ($F_{1,143}=43.15, p<.001$).

Both methods' WPM increased from sentence to sentence, however BrailleType's showed a larger improvement, with an increase from the first sentence to the last of 46.15%, against the VoiceOver's alike smaller 29.4%.

4.2 Text-Entry Accuracy

To measure accuracy, the MSD Error Rate was used, calculated as $MSD(presentedText, transcribedText) / \text{Max}(|presentedText|, |transcribedText|) * 100$ [3]. Since the data did not present a normal distribution, the non-parametric Friedman Test was used. The VoiceOver alike method, with an average MSD error rate of 14.12% against BrailleType's 8.91%, was significantly more error prone ($X^2(1)=81.94, p<.01$). This difference can be observed on Figure 4, where the majority of participants made fewer errors with BrailleType. In terms of progress along the five test sentences, results showed that when using BrailleType participants made between the first and the last phrase 5.42% less errors, while with the method identical to VoiceOver 6.7% less errors. Even though the last one had a slighter better performance increase, it still showed on average, almost twice as much errors as BrailleType (9.7% against 5.2%).

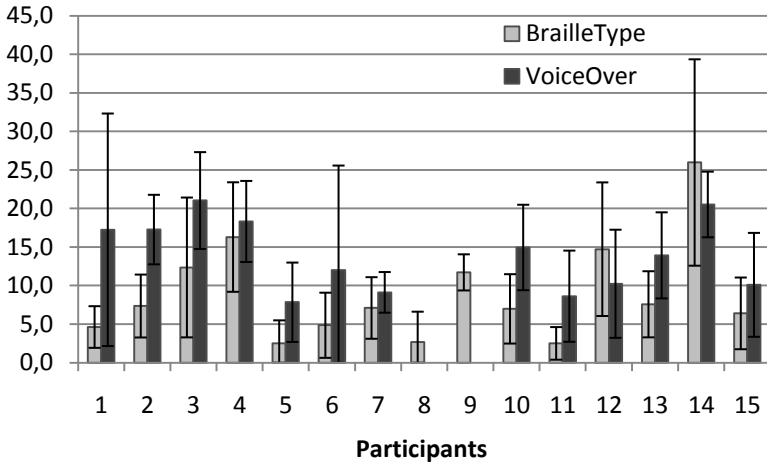


Fig. 4. MSD Error Rate (average) for each participant on each system

4.3 Users’ Feedback

Upon finishing each session, participants completed a small questionnaire about the text-entry method. This questionnaire was composed of a five-point Likert scale with four sentences and an open question about the perceived difficulties and general opinion on the method. Table 1 shows the participants’ ratings to both methods.

Table 1. Questionnaire results for each system (Median, Inter-quartile Range). The ‘*’ indicates a statistically significant response.

	Easy to comprehend. *	Easy to use. *	Quickly write text.	Would use this system.
BrailleType	5 (0)	5 (1)	3 (1)	3 (2)
VoiceOver	4 (2)	4 (2)	4 (3)	3 (1)

Users strongly agree that BrailleType is much easier to understand and to use than a system identical to VoiceOver ($X^2(1)=4.72, p<.05$ and $X^2(1)=4.92, p<.05$ respectively). Even though BrailleType was significantly slower in the tests, the participants’ opinion is that they can write almost as fast as they do on VoiceOver’s alike. However, the fact that BrailleType is a slower system weighted on the response to whether they would use the system, as the VoiceOver method was marginally preferred. In terms of perceived difficulties, the major complaint about the system identical to VoiceOver was that keys were too small and close to each other, making it hard to select and enter the intended letters. Another problem worth mentioning is that most users made involuntary errors due to split-tapping, as they would accidentally touch/rest fingers on the screen and enter unwanted letters. BrailleType was touted as a simple and easy system but slow. Most users wanted to select targets faster, which sometimes resulted in errors as they would not wait for the confirmation cue and continue without actually selecting them.

5 Conclusions and Future Work

As we move towards a future where touch screen devices threaten to replace their keypad counterparts, an effort to make them accessible to everyone is paramount. Blind people, in particular, have an added difficulty due to the absence of tactile feedback. Available solutions feature unfamiliar layouts, many targets/sub screens and/or multi-touch gestures, which although appear to be simple, can still be quite tricky to accomplish by many people.

BrailleType tries to overcome these limitations by featuring a familiar layout, the Braille cell, with few and larger targets. This approach proved to be slower than VoiceOver's approach. This does not come as a surprise since the presented method features timeouts and requires multiple inputs per character. On the other side, users commit far fewer errors with BrailleType. The large number of soft keys along with the difficulty in split-tapping, a keyboard layout with small and close targets, and errors due to involuntarily touching the screen with multiple fingers were the main reasons to this difference. These difficulties reached a pinnacle with two of the participants who were unable to write their names or a simple four-letter word, after the practice session with the help of the experimenter.

BrailleType showed promise by being an easy to comprehend, simple to use, albeit slower, text-entry method. As Braille is being less practiced due to new technologies, BrailleType could bridge these two worlds and reach out to people who would normally shy away from these devices due to unfamiliar concepts, while giving others an incentive to use Braille, as it should not be forgotten.

As future work we intend to delve further into the target selection process and test the method with progressively smaller timeout intervals or completely remove them, as we believe BrailleType can be a much faster method while keeping the errors to a minimum. A long term study with the target users will be fundamental to understand the impact of experience and the full potential of this eyes-free method.

References

1. Bonner, M., Brudvik, J., Abowd, G., Edwards, W.K.: No-Look Notes: Accessible eyes-free multi-touch text entry. In: Floréen, P., Krüger, A., Spasojevic, M. (eds.) *Pervasive Computing*. LNCS, vol. 6030, pp. 409–426. Springer, Heidelberg (2010)
2. Guerreiro, T., et al.: NavTap and BrailleTap: Non-Visual Texting Interfaces. In: RESNA 2008, Arlington, VA (2008)
3. MacKenzie, I., Tanaka-Ishii, K.: *Text entry systems: Mobility, accessibility, universality*. Morgan Kaufmann, San Francisco (2007)
4. Yfantidis, G., Evreinov, G.: Adaptive Blind Interaction Technique for Touchscreens. *Universal Access in the Information Society* 4(4), 328–337 (2006)

Potential Pricing Discrimination Due to Inaccessible Web Sites

Jonathan Lazar¹, Brian Wentz², Matthew Bogdan¹, Edrick Clowney¹,
Matthew Davis¹, Joseph Guiffo¹, Danial Gunnarsson¹, Dustin Hanks¹,
John Harris¹, Behnjay Holt¹, Mark Kitchin¹, Mark Motayne¹,
Roslin Nzokou¹, Leela Sedaghat¹, and Kathryn Stern¹

¹ Towson University, Department of Computer and Information Sciences,
Towson, MD, USA 21252
jlazar@towson.edu

² Frostburg State University, Department of Computer Science and
Information Technology, Frostburg, MD, USA 21532
bwentz@acm.org

Abstract. Although tools and design guidelines exist to make web sites accessible, a majority of web sites continue to be inaccessible. When a web site offers special prices that are available only on the web site (not the physical store), and the web site itself is inaccessible, this can lead to discriminatory pricing, where people with disabilities could end up paying higher prices than people without disabilities who can access the web site and take advantage of the online-only prices. This research examined whether 10 of the top e-commerce web sites which offer online-only price specials are accessible. The results revealed that there were multiple categories of accessibility violations found on all of the evaluated web sites.

Keywords: discrimination, web accessibility, disabilities, e-commerce.

1 Introduction

Web accessibility is the concept of making sure that web sites can work properly for users with disabilities that are using alternative input or output devices, such as screen readers or adaptive keyboards. Typically, for a web site to be accessible, it must follow a set of design guidelines, such as the web guidelines from Section 508 (U.S. Law), or the Web Content Accessibility Guidelines (WCAG) from the Web Accessibility Initiative (WAI). Other countries also have legislation that relates to web site accessibility, including the Equality Act in the U.K. [6], Act on Equal Opportunities for Disabled Persons in Germany [4], and the Disability Discrimination Act in Australia [1], to name a few.

Although tools and design guidelines exist to make web sites accessible, a majority of web sites continue to be inaccessible. Various authors have documented how U.S. federal and state government web sites, university web sites, airline web sites, e-commerce sites, and employment web sites continue to be inaccessible

[5],[7],[8],[8],[10]. However, there is an additional problem: web site inaccessibility often leads to other unwanted societal effects, such as pricing discrimination. A previous study documented that, when airline web sites are inaccessible, people with disabilities must use the airline call center, which can lead to the individual paying a higher airfare (even though that is against the law) [11].

For many stores, there are two components to the enterprise: the physical, brick and mortar store, and the online, e-commerce site. These two components are often tightly integrated (for instance, stores such as Wal-Mart allow a customer to order from the web site and have an item shipped to a local store), but many e-commerce sites offer special web-only prices which are not available in the physical store or over the phone. When a web site offers special prices that are available only on the web site, and the web site itself is inaccessible, this can lead to discriminatory pricing.

2 Research Methodology

There were two components to this research: stage one was determining which large stores have both physical and online components and offer special web-only prices, and stage two was evaluating those web sites for accessibility. The 50 largest e-commerce sites in the U.S. according to [2] were narrowed to 41 sites which had both physical stores and e-commerce sites (the other nine e-commerce sites, such as Amazon.com and Zappo's, had no physical counterpart). Next, those 41 web sites were examined to determine which sites had web-only special prices using the following five-step process:

1. Examined a few items, to see if there was a separate online-only price.
2. Searched for a "hot deals" (or something similar) section on the web site, and checked for online-only deals. Checked if there are any flyers or deals located only in a local area (if a zip code is required).
3. Searched the e-commerce site's search engine using the phrase "online only."
4. Searched using Google advanced search, set to the domain of the e-commerce site, and the exact phrase "online only" (because many web site search engines are actually product search engines, not keyword search engines).
5. Signed up for an email mailing list for specials, and monitored for a week to determine if any web-only specials were sent out.

After evaluating for online-only prices or specials, it was determined that the 10 largest stores that had online-only prices (in order, starting from the largest) were: Staples, Office Depot, OfficeMax, Sears, BestBuy, Costco, Victoria's Secret, Macy's, Gap Direct, and Neiman-Marcus. To evaluate the accessibility of the web sites during stage two, expert inspections were conducted (with multiple, independent evaluators for each web site) using a screen reader, since this is considered to be the most effective form of accessibility evaluation for compliance with standards [12] and most accurate when multiple evaluators inspect the same web site [10].

The expert evaluations were conducted during November and December 2010 by utilizing the web site accessibility standards of Section 508 (1194.22) of the U.S. Rehabilitation Act, which involve 16 guidelines, which are referred to as paragraphs

“a” through “p” [14]. The expert evaluation process was guided by the “Absolute Minimum Accessibility Inspection” which has been used effectively in other accessibility evaluations [10]. Each “paragraph” of the Section 508 guidelines is considered to be weighted equally, and the existence of a paragraph violation for a web site was recorded. Each web site was individually evaluated by a minimum of four individuals without disabilities (due to the total number of evaluators involved, some web sites received five evaluations). All evaluators then met, discussed their individual evaluations, and compiled one meta evaluation, which typically has a higher level of validity than only one evaluation [13]. Table 1 is a summary of the Section 508 guidelines with the associated “paragraph” letters.

Table 1. Summary description of the 16 paragraphs of the U.S. Section 508 web guidelines

<i>Paragraph</i>	<i>Summary Description</i>
A	Text Equivalent (have a text equivalent for any graphical elements)
B	Synchronized Equivalent Alternatives (have captioned video, transcripts of any audio, or other alternatives for multimedia)
C	Use of Color (color should not be used as the only method for identifying elements of the web page or any data)
D	Organization (style sheets are encouraged, but users should still be able to utilize a web page when style sheets are turned off)
E, F	Redundant Text Links on Server-Side Image Map and (f) Client-Side Image Maps (redundant clickable links for server-side image maps, and accessible client-side image maps are preferred)
G, H	Row and Column Headers (use appropriate headers and markup to allow easy navigation of a table)
I	Frames (title all frames and label all frames for easy identification and navigation, e.g., use “navigation” “main content” and “search” rather than “top” or “bottom”)
J	Screen Flicker Frequency (limit or eliminate the use of flickering, which can provoke seizures)
K	Text-Only Page Default (if a web page cannot be made accessible, provide an equivalent text-only page, and make sure it is kept up to date)
L	Scripting Languages (make sure that equivalents for any non-accessible scripting are included, e.g., for those who are not using pointing devices)
M	Linked Plug-In or Applet (if any plug-ins are required, make sure to provide a link to an accessible version of the plug-in)
N	Online Electronic Forms (all forms must be properly labeled and accessible)
O	Method to Skip Repetitive Navigation Links (all web pages should have a link which allows a user to skip directly to the main content, bypassing any site navigation information)
P	Alerts on Timed Responses (if any page responses are timed, the user should be given the opportunity to indicate that more time is needed)

3 Results

The accessibility evaluations of these web sites revealed that all 10 web sites violated at least one paragraph of the Section 508 guidelines, with an average of three

paragraphs violated per web site. The variety of violations ranged from OfficeMax violating only paragraph a (alternate text equivalent) to Costco violating five paragraphs (a, d, h, i, and o). Staples violated paragraphs a, b, and m with buttons and images that did not have alternate text and flash content without an alternative. An example of the impact this could have on pricing was that the Staples “Cyber Monday” links were not accessible. OfficeDepot also violated the same three paragraphs for similar reasons, and a promotional video lacked transcripts or captioning. Office Max violated only paragraph a (lack of text equivalent for some of the graphical elements on the site). Sears violated paragraphs a, b, and o with lack of alternate text, no method of skipping repetitive navigational links and the lack of transcripts or captioning for a promotional deals video.

BestBuy violated paragraphs a, b, f, and m with the lack of alternate text, a flash video with no alternative way to reach or read the content, and no links to the required plug-in. An example of the impact that this could have on pricing discrimination is the “deals” section that cannot be accessed without the use of a mouse on the BestBuy web site as illustrated in Figure 1.

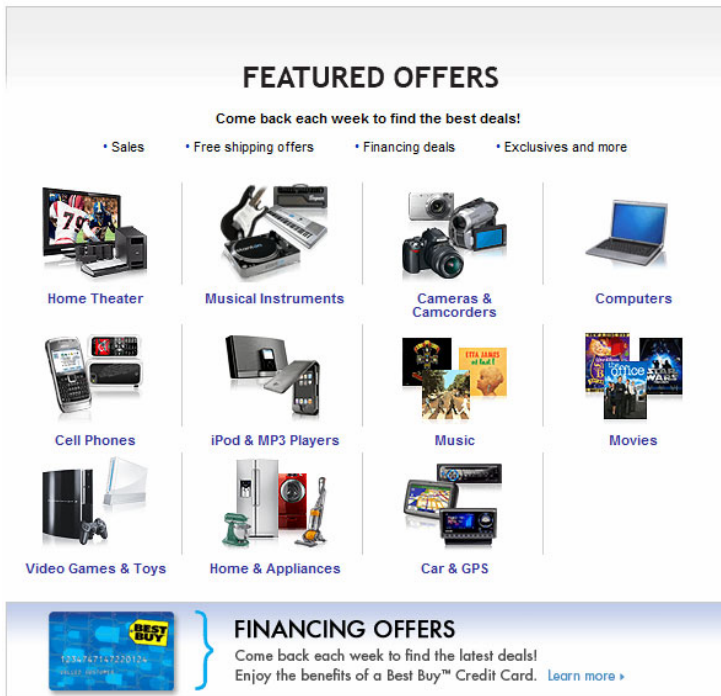


Fig. 1. Screenshot of the deals section on the BestBuy web site which cannot be accessed by the keyboard alone

Costco had violations in five paragraphs (a, d, h, i, and o), and those violations included the lack of functional skip navigation links, poor alternate text, lack of functionality without the style sheet, and frames and tables that were not labeled or poorly

labeled. Victoria’s Secret violated paragraphs a and o, with lacking alternate text and a method to skip repetitive navigational links. The Macy’s web site violated paragraphs a, b, and l with lack of alternate text, inaccessible videos, and a script to display a promotional code that was not accessible with a screen reader. GapDirect violated paragraphs a, l, m, and n with lack of alternate text, inaccessible scripting, no link for the flash plug-in, and unlabeled form fields. Neiman-Marcus violated paragraphs a, n, and o with lack of alternate text on images and buttons, forms with incorrect labels, and no link to skip repetitive navigational links. Table 2 shows the results of which web sites violated which paragraphs of the Section 508 guidelines.

Table 2. Results of Violations of the Section 508 Guidelines

Summary of Section 508 Paragraph Violations	Staples	OfficeDepot	Office Max	Sears	BestBuy	Costco	Victorias Secret	Macy’s	Gap Direct	Neiman-Marcus
a) Alternate text for images/other visuals	x	x	x	x	x	x	x	x	x	x
b) Synch. Multimedia	x	x		x	x			x		
c) Meaning through color also avail. w/o										
d) Readable w/o CSS						x				
e,f) Server/Client-side image maps					x					
g,h) Table headers/markup						x				
i) Frames have labels						x				
j) No blinking or flashing										
k) Text page if needed										
l) Scripting languages								x	x	
m) Applets/Plug-ins	x	x			x				x	
n) Forms									x	x
o) Skip navigation				x		x	x			x
p) Timed response										
Total Categories Violated	3	3	1	3	4	5	2	3	4	3

4 Conclusion

Companies must pay careful attention to the accessibility of their interfaces, particularly so that all individuals are provided equal access to all content. If people with disabilities cannot access pricing “deals” or “specials” on e-commerce web sites, it could lead to pricing discrimination. Discrimination against an individual on the basis of disability is clearly against the law in many countries. In the U.S., the 2007 court case involving Target.com [3] illustrated the necessity for businesses to pay closer attention to web site accessibility.

Individuals involved in designing or maintaining these interfaces must carefully design to standards, such as WCAG. WCAG (Web Content Accessibility Guidelines) form the basis for web accessibility policies throughout the world, including Section 508 in the U.S. One example of how the accessibility violations discovered in this study conflict with WCAG is seen in WCAG 2.0 Principle 2, which specifies that interfaces must be navigable through a keyboard interface [15].

Some basic recommendations that can significantly improve the accessibility of a web site can be derived from either WCAG or other regulations, such as Section 508 in the U.S. (recall Table 1). WCAG 2.0 summarizes its guidelines by an interface being perceivable, operable, understandable, and robust [15]. Perceivable means that an interface must provide alternatives for the types of media that are presented, whether inherently visual, auditory, or haptic. Operable means that all users can read and use the content, even from a keyboard alone. It also means that users should have enough time to read content, be easily able to know where they are, and be confident that the design of an interface will not inherently cause a seizure. Understandable means that content should be readable and easy to understand, have predictable operation, and assist users with avoiding and correcting mistakes. Robust means that an interface should be able to be accessed regardless of the technology used to access it, including assistive technologies [15].

Policy makers and those responsible for enforcing current laws (including civil rights legislation) should be aware of the impact that web inaccessibility can have on the civil rights of individuals with disabilities. Following guidelines such as WCAG and performing regular evaluations on the accessibility and usability of an interface, involving users with disabilities, accessibility experts, and automated evaluation tools will help to prevent possible discrimination problems, such as the ones discussed in this study. More research needs to be done on how inaccessible web sites can lead to unwanted and possibly illegal actions such as pricing discrimination, employment discrimination, and societal exclusion.

References

1. Australian Human Rights Commission: Disability Rights, http://www.hreoc.gov.au/disability_rights/
2. Davis, D. (ed.): *Internet Retailer: 2009 Top 500 Guide*, Vertical Web Media, Chicago (2009)
3. Frank, J.: *Web Accessibility for the Blind: Corporate Social Responsibility or Litigation Avoidance?* In: 41st Hawaii International Conference on System Sciences, pp. 1–8 (2008)
4. Federal Ministry on Labor and Social Affairs: *Disability Policy*,

- http://www.bmas.de/portal/45136/disability__policy.html
5. Gladstone, K., Rundle, C., Alexander, T.: Accessibility and Usability of eCommerce Systems. In: Miesenberger, K., Klaus, J., Zagler, W.L. (eds.) ICCHP 2002. LNCS, vol. 2398, pp. 11–18. Springer, Heidelberg (2002)
 6. Government Equalities Office: Equality Act 2010, http://www.equalities.gov.uk/equality_bill.asp
 7. Hull, L.: Accessibility: it's not just for disabilities any more. *Interactions* 11(2), 36–41 (2004)
 8. Jaeger, P.T.: Assessing Section 508 compliance on federal e-government Web sites: A multi-method, user-centered evaluation of accessibility for persons with disabilities. *Government Information Quarterly* 23(2), 169–190 (2006)
 9. Kane, S., Shulman, J., Shockley, T., Ladner, R.: A web accessibility report card for top international university web sites. In: 2007 International Cross-Disciplinary Conference on Web Accessibility (W4A), pp. 148–156. ACM, New York (2007)
 10. Lazar, J., Beavan, P., Brown, J., Coffey, D., Nolf, B., Poole, R., Turk, R., Waith, V., Wall, T., Weber, K., Wenger, B.: Investigating the Accessibility of State Government Web Sites in Maryland. In: Langdon, P., Clarkson, P., Robinson, P. (eds.) *Designing Inclusive Interactions*, pp. 69–78. Springer, London (2010)
 11. Lazar, J., Jaeger, P.T., Adams, A., Angelozzi, A., Manohar, J., Marciniak, J., Murphy, J., Norasteh, P., Olsen, C., Poneres, E., Scott, T., Vaidya, N., Walsh, J.: Up in the Air: Are Airlines Following the New DOT Rules on Equal Pricing for People with Disabilities When Websites are Inaccessible? *Government Information Quarterly* 27(4), 329–336 (2010)
 12. Mankoff, J., Fait, H., Tran, T.: Is your web page accessible?: a comparative study of methods for assessing web page accessibility for the blind. In: *SIGCHI Conference on Human Factors in Computing Systems*, pp. 41–50 (2005)
 13. Nielson, J., Mack, R. (eds.): *Usability Inspection Methods*. John Wiley and Sons, New York (1994)
 14. U.S. Government: Section 508, <http://www.section508.gov>
 15. WC3. Web Content Accessibility Guidelines (WCAG) 2.0 (2008), <http://www.w3.org/TR/WCAG/>

Measuring Immersion and Affect in a Brain-Computer Interface Game

Gido Hakvoort, Hayrettin Gürkök, Danny Plass-Oude Bos,
Michel Obbink, and Mannes Poel

University of Twente, Faculty EEMCS,
P.O. Box 217, 7500 AE, Enschede,
The Netherlands
{gido.hakvoort,mobbink}@gmail.com,
{h.gurkok,m.poel}@utwente.nl, d.oudebos@cs.utwente.nl

Abstract. Brain-computer interfaces (BCIs) have widely been used in medical applications, to facilitate making selections. However, whether they are suitable for recreational applications is unclear as they have rarely been evaluated for user experience. As the scope of the BCI applications is expanding from medical to recreational use, the expectations of BCIs are also changing. Although the performance of BCIs is still important, finding suitable BCI modalities and investigating their influence on user experience demand more and more attention. In this study a BCI selection method and a comparable non-BCI selection method were integrated into a computer game to evaluate user experience in terms of immersion and affect. An experiment with seventeen participants showed that the BCI selection method was more immersive and positively affective than the non-BCI selection method. Participants also seemed to be more indulgent towards the BCI selection method.

Keywords: Brain-computer interfaces, affective computing, immersion, games.

1 Introduction

A brain-computer interface (BCI) can be described as a communication link between the brain and the machine. In a BCI system, signals from the brain are analyzed to determine the user's state of mind or intentions, which in turn can be translated into actions [9]. BCI systems have been applied for medical use to help disabled users by giving back mobility [8] and breaking the isolation of people with physiological disorders [7,11].

As successful applications of BCIs become news and commercial BCI input devices become publicly available, BCIs are finding their way into recreational use. However, as the scope of BCI applications is expanding from medical to recreational use, the expectations for BCIs are also changing. Currently they are unable to meet the high performance and accuracy of existing input modalities such as mouse and keyboard, and are therefore unfit as replacement. Instead they should be seen as

separate modalities which can be used beside, or in combination with, existing input modalities [17]. However, using BCI as input modality still comes with many challenges. Where increasing the performance of BCIs has always been an important goal for medical studies, the way they are applied as modalities and the influence they have on the user experience is becoming more and more important for recreational BCI applications.

Whether BCIs are suitable for recreational applications is as yet unclear as they have rarely been evaluated for user experience. They may turn out to be valuable additions for recreational applications such as games which are developed to be challenging and enjoyable. Moreover the inaccuracy of BCIs can become a challenging factor in games. As gamers love working with new technologies, are capable of adapting quickly to a new environment and are used to the concept that games have to be mastered [16], they may be more indulgent towards BCI modalities.

The purpose of this study was to evaluate a BCI system for user experience in the terms of affect and immersion. As making selections is an important aspect in many games and BCIs are frequently used to make selections, a BCI selection method was used in this study. One of the most frequently used brain signals in BCI systems to make selections is the steady-state visually evoked potential (SSVEP) [1]. In most cases SSVEPs are triggered by presenting a modulated visual stimulus with a periodic signal, usually at frequencies above 5 Hz, to a user. The periodic signal of the stimulus can then be traced back in the measured brain signals which are mostly recorded from the occipital region of the scalp [23]. The power of an SSVEP only covers a narrow bandwidth matching that of the stimulus [15]. This makes them relatively easy to detect, which is why the BCI selection method used in this study was based on SSVEP. The BCI selection method and a comparable non-BCI selection method were integrated into a computer game where both introduced a challenge factor. As the BCI selection method would be able to directly translate the user intentions into in-game actions, it was expected that it would enrich the user experience in terms of immersion and affect.

How immersion and affect can be influenced by input modalities will be explained in the background in section 2. The BCI and non-BCI selection methods and how they were integrated in a game will be described in section 3. In section 4 the experimental setup and how both selection methods were evaluated will be explained. After this, the results of the experiment will be reported, followed by the discussion and the conclusion.

2 Background

2.1 Immersion

Immersion has meaning in various contexts, such as while reading a book, watching a movie or playing games. Whether the term is used consistently in these contexts is

unclear. However, for playing games there seems to be a shared concept of immersion among gamers [4]. Immersion in games is often accompanied by high levels of concentration, losing track of time and empathy.

In a study by Brown *et al.* [4] an attempt was made to define immersion within games. In their study they examined the concept of immersion experienced by gamers. Their results indicate that immersion is not just a static experience, but has different levels of involvement. They defined three levels of involvement: engagement, engrossment and total immersion. They also state that to reach a certain level of immersion, a number of barriers must be crossed. Some of these barriers are related to human characteristics, such as personal preferences, empathy and the will to invest time and attention. Others are related to the construction of the game such as the graphics, a plot and atmosphere. However, to reach a state of total immersion an important barrier to take is related to the controls. Through them gamers translate their intentions into in-game actions [19] and virtually controls should become invisible to the gamer. As intentions originate from the brain, users first need to translate their intentions into real world actions to handle the controls. Even if these real world actions become virtually invisible for the user, they still need to be performed. Using a BCI to detect intentions may allow them to be translated directly into in-game actions, making the real world actions redundant.

In a study by Jennett *et al.* [12] immersion in games was further investigated. They identified five factors of immersion: cognitive involvement, real world dissociation, emotional involvement, challenge and control. As in the study of Brown *et al.*, some factors were related to human characteristics (cognitive involvement, real world dissociation and emotional involvement) and others were related to the construction of the game (challenge and control). To measure these factors, as well as the total immersion, they developed a questionnaire which was also used in this study to measure the immersion while using the BCI and non-BCI selection methods.

2.2 Affect

Affect can be referred to as experiencing emotions and has some overlap with immersion [12]. It has a large impact on how well users are able to perform tasks and how they respond to possible usability problems. According to Norman [18], a more positive affect causes users to be more indulgent towards minor usability problems. Although there are many dimensions associated with affect, according to Picard [20] the three most commonly used dimensions of emotion are valence, arousal and dominance. Picard also notes that the valence and arousal dimensions are critical in recreational applications.

Bradley *et al.* [3] developed a questionnaire, the the self-assessment manikin (SAM), to measure emotional responses in these three dimensions. In Fig. 1 some emotions associated with valence and arousal can be seen. Integrating the selection methods into an enjoyable, challenging and task oriented environment, such as a game, should result in a more positive affect in terms of valence and arousal which will aid users to overcome the inaccuracy of the selection methods.

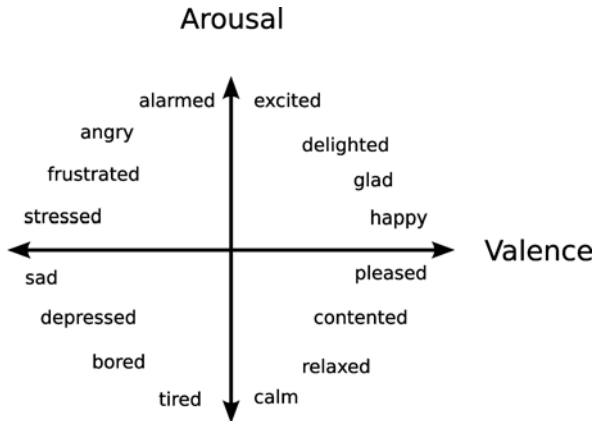


Fig. 1. Emotions in the valence and arousal space

3 Selection Methods

To measure the effects of the BCI and non-BCI selection methods on the user experience in terms of immersion and affect, the selection methods were integrated into a simple game. The game, called *Mind the Sheep!*, offered the players an enjoyable, challenging and task oriented environment.

3.1 *Mind the Sheep!*

Mind the Sheep! consists of a playground representing a meadow on which a few obstacles, such as fences, gates or trees are placed. There are three dogs in the playground which are controlled by the player. Depending on the objectives of a specific level, the playground can be populated with a number of sheep or collectibles. A screenshot of the game can be seen in Fig. 2 with ten white sheep and three black dogs. The goal of the game is to gather all sheep into a pen or gather all collectibles as quickly as possible using the three dogs. Players can use one of the selection methods to select one of the three dog.

To move a dog, players point at any location on the playground with the mouse and start the current selection method by pressing the left mouse button. As long as they hold down the mouse button the selection method continues to be operative thereby increasing the accuracy of the selection method. Releasing the mouse button stops the selection method and one of the dogs is moved to the location indicated by the player. When players indicate a position unreachable for the dogs, the instruction is ignored. As the accuracy of the selection methods increase over time, there is a trade-off between accuracy and speed. It also requires users to multi-task as they need to concentrate on a dog and keep an eye on the sheep at the same time.

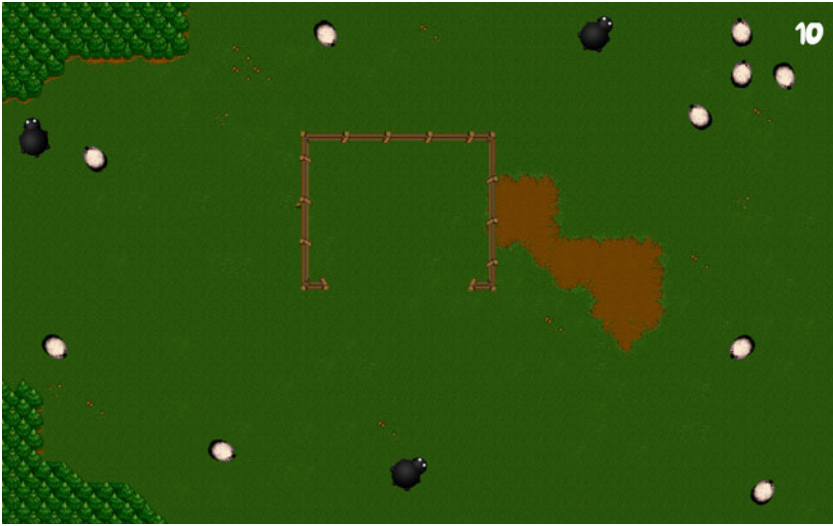


Fig. 2. Screenshot of Mind the Sheep! with three black dogs and ten white sheep. The pen is located in the center.

The dogs use a simple A* search based path finding algorithm to move to a specific location on the map. By default, sheep will walk around and graze randomly. However, when a dog approaches, they will tend to flock and move away from the dog allowing them to be herded in a desired direction. The flocking behavior is introduced by using the boids algorithm [22]. By positioning the dogs at strategic locations on the map a flock of sheep can be directed into the pen.

3.2 BCI Selection

To use SSVEPs for making selections, stimuli emitting unique periodic signals should be presented to a user simultaneously. When a user focuses on one of the stimuli, the periodic signal of that stimulus can be traced back in the user's brain signals. As each stimulus represents a single target, the corresponding target can then be determined as the selected one. As users would not want to continuously make selections, they were able to start and stop this selection method thereby controlling the presence of the stimuli. The size of the stimuli and the frequency of their periodic signals have an influence on the how well they are reflected in user's brain signals. Therefore, these properties were determined in a pre-experiment study.

In the pre-experiment study a small white cross was placed at the center of an LCD monitor on a black background (Fig. 3). During the study participants were exposed to different types of trials, in which the presented stimulus varied in size and frequency. The detailed setup of the pre-experiment study was described in [10].

Based on the work of Volosyak *et al.* [24], 7 different frequencies (6 Hz, 6.67 Hz, 7.50 Hz, 8.57 Hz, 10.00 Hz, 12.00 Hz and 15.00 Hz) were used. Their study showed that frequencies which are integer factors of the refresh rate of an LCD monitor are more suitable for presenting stimuli. Using the factor of the refresh rate produces a

more stable stimulus frequency because only whole frames are visible on the LCD screen. The diameters of the stimuli were set to 2 and 3 centimeters, which is consistent with the work in the literature [5,14].

In each trial a simple stimulus, a blinking white circle, appeared at the location of the cross. The participant focussed for 4 seconds on the stimulus. Between trials, participants had 6 seconds rest to relax their vision. All trials were presented 25 times and were placed in a random order prior to the study, thus for each participant, 25 segments of 4 seconds of data were recorded for each different trial.

The results of the pre-experiment study with 7 participants indicated that a good set of frequencies is 10 Hz, 12 Hz and 15 Hz with a diameter of 3 centimeters. Without any training, the frequencies in this set were classified correctly with an average recall of 78.1 % ($\sigma = 18.5$) using a CCA-based detection method which will be further explained in section 4.2.

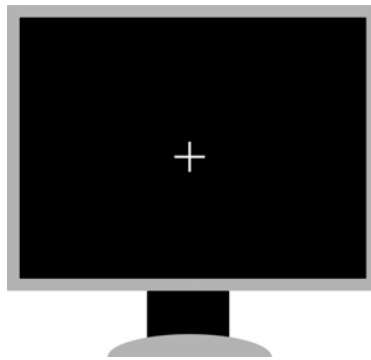


Fig. 3. Screen layout of the monitor during the pre-experiment study

3.3 Non-BCI Selection

The BCI selection method introduces a challenge factor when a selection is being made as users need to concentrate on a stimulus to make an accurate selection. The main challenge is to gather enough data to be able to make a good detection, which is directly related to the time the selection method is operative. Although it would be much easier for users to make a selection by pointing and clicking with a mouse, this would offer little challenge and give the non-BCI selection method a great advantage when compared with the BCI selection method. To pose a similar challenge in the non-BCI selection method, it was modelled to be similar to the BCI selection method in terms of the time required for an accurate selection.

Similar to the BCI selection method users were able to start and stop the non-BCI selection method. When operative, the dogs were highlighted one at a time with an increasing highlight period. The diameter of the stimuli were kept the same as in the BCI selection method, 3 centimeters. When users stopped the selection method, the current highlighted dog was selected.

To make an accurate selection, users had to react in time when the dog they wanted to select was highlighted. With a highlight period of 250 ms users should have enough

time to react as this is close to the average human reaction time [13]. The highlight period should reach 250 ms only by the time that is equal to the time needed to make a selection using the BCI selection method, which is around 2.5 seconds [14]. Therefore, the highlight period started with 100 ms and was increased 5% after every highlight, with a maximum highlight period of 500 ms.

4 Methods

4.1 Experimental Setup

The experiment held in this study consisted of two different sessions. In each session a participant used one of the selection methods (BCI or non-BCI) while playing *Mind the Sheep!*. Each participant used both selection methods. However, a counterbalanced measures design was used to avoid confounding variables such as learning a strategy to play the game.

Each session was divided into three trials, a familiarity trial, an easy trial and a difficult trial. In the familiarity trial participants could get used to the selection method by selecting and moving the dogs. During this trial, participants had to collect 10 objects which were placed across the playground. Next, in the easy trial, participants had to pen a small flock of 5 sheep using the dogs. During this trial two pens were placed on the playground, one on the left and one on the right of the screen to make the task easier for the participants. Finally, in the difficult trial, participants had to gather 10 sheep, which were more scattered across the playground, into one pen that was placed in the center of the playground. Between the two sessions, participants were given a break of ten minutes.

The layout of the playgrounds across the trials were kept the same to ensure no playground was more difficult for one of the selection methods. However to ensure that participants did not create an optimal strategy for a specific trial, the positions of the dogs, collectible objects and sheep were altered for the different selection methods.

A timeout was set for each trial for the participants to finish the level by collecting all objects or gathering all sheep into a pen. Participants had 3 minutes, 5 minutes and 10 minutes for the familiarity, easy and difficult trials respectively. Since immersion in games is often accompanied by losing track of time, the time left was not visible for the participants. Otherwise it could have influenced their perception of the elapsed time.

The game ran on a PC which was connected to a projector¹. The projector was mounted on the ceiling and projected the game on a screen approximately 3 meters away from a participant. The sizes of the stimuli were scaled proportionally with the increased distance to the screen. To make sure the frequencies for the SSVEP based BCI selection method were correctly presented by the projector they were checked with a light sensor.

Participants sat on a chair behind a table and by using the mouse on the table they were able to start and stop the selection methods. The data acquisition ran on a separate PC and sent the raw electroencephalography (EEG) data to the game PC.

¹ Mitsubishi WD510U, 93.3", 60Hz, 1360x768.

4.2 SSVEP Detection

Prior to the experiment, 32 electrodes were placed according to the international 10-20 system [21]. For the EEG data acquisition a BioSemi ActiveTwo system² was used. During each SSVEP selection EEG data was used from eight parietal-occipital electrodes (i.e. *Pz*, *P3*, *P4*, *PO3*, *PO4*, *Oz*, *O1*, *O2*). The EEG data was re-referenced to the common average reference (CAR) [6] of all 32 electrodes, after which a detection method was used to determine which frequency the participant was focusing on.

There are various methods to detect the presence of SSVEP. One of the most popular and widely used detection methods is power spectral density analysis (PSDA) where a fast Fourier transform (FFT) is used to estimate the power spectral density (PSD) of a time window of the EEG signal. The magnitude of each stimulation frequency can then be used for classification. A relatively new approach is using canonical correlation analysis (CCA) where sets of reference signals are constructed for each one of the stimulation frequencies. Each set contains the *sine* and *cosine* for the first, second and third harmonic of the stimulation frequency. The re-referenced EEG data and each set with reference signals are used as input for the CCA. CCA tries to find pairs of linear transformations for the two sets such that when the transformations are applied the resulting sets have a maximal correlation. The stimulation frequency with the highest maximum correlation is classified as the frequency the participant was focusing on.

As CCA-based detection methods have some improvements and advantages compared to PSDA-based detection methods, such as better signal-to-noise ratio (SNR), lower inter-subject variability and the possibility of using harmonic frequencies [2], a CCA-based detection method was used in this study.

4.3 Questionnaires and Data Acquisition

To measure the affective reaction of the participants while playing the game, they were requested to fill in a SAM [3] after each trial. It was expected that if participants became more frustrated by using a selection method, this would result in higher arousal and lower valence scores. The dominance would also be higher if participants had the feeling that a selection method was working properly and that they were in control.

After each session participants filled in a questionnaire on immersion. The questionnaire by Jennett *et al.* [12] was used for this. It contains 31 questions and is designed to measure the total immersion as well as five different factors of immersion (cognitive involvement, emotional involvement, real world dissociation, challenge and control). Although valence, arousal and dominance would probably differ between selection methods for which participants have an aversion, this does not necessarily mean participants would not become immersed in the game.

Furthermore some game statistics, such as the number of times a dog was selected, the total number of selections, the average selection duration and the time participants needed to finish a trial were collected while participants played the game. Ideally participants would use all the dogs as this would make it easier to gather the sheep. The number of times a particular dog was selected would be a good indication of how well the selection methods performed for the participants.

² BioSemi, Amsterdam, The Netherlands.

At the end of the experiment participants were asked which selection method they would like to use if they were given the opportunity to play the game again. This should have given a good indication of which selection method the participant preferred.

After the experiments, scores were obtained for the total immersion, the five immersion factors, the SAM questionnaire and the game statistics. The results from the SAM questionnaire were averaged over the three trials to get an average score for the valence, arousal and dominance scales. As each participant used both selection methods, the scores were compared using a Wilcoxon signed-rank test to determine if there was a significant difference between the two selection methods.

4.4 Participants

Seventeen participants (7 female and 10 male), aged between 17 and 37 ($\mu = 22.00$, $\sigma = 4.74$) participated in the experiment. All participants except for one had normal or corrected-to-normal vision and described themselves as daily computer users. Although eight participants had at least one-time experience with EEG, fourteen participants had no experience with BCIs. Before the experiment, all participants signed an informed consent form and they were paid according to institution's regulations.

5 Results

When the 17 participants were asked which selection method they would choose if they were given the opportunity to play the game again, 5 of them chose the non-BCI selection method and 12 of them chose the BCI selection method. This is a first indication that participants prefer the BCI selection method over the non-BCI selection method. A detailed insight might be provided by the scores for immersion, the SAM questionnaire and the game statistics, which are described below.

5.1 Immersion

Based on the immersion questionnaire the total immersion score was calculated for both selection methods. On average the participants rated the BCI selection method ($\mu = 160$, $\sigma = 14.55$) higher than the non-BCI selection method ($\mu = 144$, $\sigma = 21.89$). This difference was significant ($Z = -2.155$, $p = 0.031$).

The five immersion factors were also analyzed and the scores, averaged over participants, are shown in Table 1. The scores for all factors, except for the challenge factor, are higher for the BCI selection method. The five immersion factors were examined for

Table 1. Scores of the five immersion factors, averaged over participants. Values are represented as $\mu(\sigma)$ with * indicating a significant difference with $p < 0.05$.

	non-BCI	BCI
Cognitive *	53.12 (9.01)	58.47 (5.36)
Dissociation *	25.18 (5.64)	28.06 (4.79)
Emotional *	54.35 (8.54)	59.88 (5.95)
Challenge	20.88 (2.89)	20.71 (2.34)
Control *	32.06 (5.77)	37.06 (5.08)

significant differences between the BCI and non-BCI selection methods. There are significant differences for the cognitive involvement factor ($Z = -2.219$, $p = 0.026$), the real world dissociation factor ($Z = -1.992$, $p = 0.046$), the emotional involvement factor ($Z = -2.013$, $p = 0.044$) and the control factor ($Z = -2.310$, $p = 0.021$). For the challenge factor no significant difference ($Z = -0.476$, $p = 0.634$) was found.

5.2 Affect

Based on the SAM questionnaire the total SAM scores were calculated for both selection methods and the average results are shown in Table 2.

Table 2. Average SAM scores, the values are represented as $\mu(\sigma)$ with * indicating a significant difference of $p < 0.05$

	non-BCI	BCI
Valence*	6.37 (1.55)	7.08 (1.31)
Arousal	4.82 (2.31)	4.53 (2.37)
Dominance	5.65 (2.46)	6.08 (1.64)

For the valence scale the difference was significant ($Z = -2.012$, $p = 0.044$), however, no significant difference was found for the arousal ($Z = -0.315$, $p = 0.752$) or dominance ($Z = -0.403$, $p = 0.687$) scores.

5.3 Game Statistics

The game statistics which were collected during the experiments are shown in Tables 3 and 4. They show the number of selections and the average selection time. For the number of selections (Table 3) there are significant differences for the easy task trial ($Z = -2.202$, $p = 0.028$) and the difficult task trial ($Z = -2.107$, $p = 0.035$). However, no significant difference was found for the familiarity trial ($Z = -1.866$, $p = 0.062$).

Table 3. Average number of selections for each trial, presented as $\mu(\sigma)$ with * indicating a significant difference of $p < 0.05$

	non-BCI	BCI
Familiarity	12.53 (4.87)	10.47 (2.67)
Easy*	43.00 (26.27)	27.88 (7.86)
Difficult*	98.88 (65.90)	74.12 (30.28)

Table 4. Average selection time (in seconds) for each trial, presented as $\mu(\sigma)$

	non-BCI	BCI
Familiarity	1.92 (0.86)	2.49 (1.71)
Easy	1.96 (1.04)	2.64 (2.63)
Difficult	1.95 (1.31)	1.71 (0.85)

Although no significant differences were found for the average selection time (Table 4) between the BCI and non-BCI selection methods for any of the three trials, it can be seen that for the familiarity and easy task trials the average selection times for the BCI selection were around 0.5 seconds higher. However, for the difficult task trial the average selection time was lower.

6 Discussion

Most participants indicated that they preferred playing the game with the BCI selection method. This seems to be supported by the results of the immersion questionnaire. The total immersion score was significantly higher for the BCI selection method, indicating that participants were more immersed than when using the non-BCI selection method. For the BCI selection method participants only had to make their intentions clear to the BCI system by concentrating on the dog they wished to selected. Besides starting and stopping the selection method, there were no other actions required. However, for the non-BCI selection method, participants still had to translate their intentions into an action, stopping the selection method at the correct time. As participants were able to translate their intentions directly into in-game actions while using the BCI selection method, they might have become more easily immersed.

Further inspection of the five immersion factors showed that all factors, except for the challenge factor, were significantly higher for the BCI selection method. The questions related to the challenge factor were about the game itself. As the averages of the challenge factor were almost equal for the two selection methods, they indicate that participants did not find the game more challenging using one of the two selection methods. For using the selection methods, the control factor might be a better indicator. As the control factor was significantly lower for the non-BCI selection method, it indicates that participants had more trouble using the non-BCI selection.

The results of the SAM questionnaire indicate that participants were more content using the BCI selection method. The results of the cognitive involvement and real world dissociation factors were also significantly higher for the BCI selection method. This could have been caused by the fact that participants had to concentrate on a dog while using the BCI selection method and did not have to translate their intentions into real world actions.

Some participants developed a strategy to deal with the trade-off between accuracy and speed. When they wanted to be accurate in their selections, they waited long enough to make an accurate selection. However, when they wanted to be quick, they moved all three dogs as one by pressing rapidly at a location on the playground. Although this behavior was observed for both selection methods, for the BCI selection method participants appeared to switch between precise and quick whenever they wished. However, for the non-BCI selection method they appeared to prefer only making quick selections, explaining the significantly higher number of selection for the non-BCI selection method.

For the BCI selection method the number of selections was lower and, although not significant, the average selection time was higher. Participants were also more

immersed and content. Apparently, participants accepted that for a good SSVEP detection they had to wait a couple of seconds. However, for the non-BCI selection method they understood that it was related to their own reaction speed and suddenly they appeared to be in a hurry. Although the non-BCI selection method was modeled to be similar to the BCI selection method, it might have had an effect on the preference of the participants, as they did not want to wait to make a selection. The non-BCI selection method did introduce the same challenge and required participants to multi-task similarly to the BCI selection method. However, participants seemed to be more indulgent towards the BCI selection method than towards the non-BCI selection method. This could be caused by the curiosity of participants for the BCI selection method or the self overestimation by participants while using the non-BCI selection method.

7 Conclusion

In this study a BCI system was compared to a non-BCI system to evaluate the user experience in terms of immersion and affect. For the BCI system a selection method based on SSVEP was integrated into a game, introducing a challenge factor. A comparable non-BCI selection method based on time was also implemented into the game, introducing an equal challenge. Seventeen participants played the game with both selection methods in three trials. They rated each selection method on immersion and affect.

The results show that the BCI selection method was found to be more immersive and positively affective than the non-BCI selection method. While using the BCI selection method participants were able to directly translate their intentions into in-game actions, which made it easier for them to become more immersed. In this case, the user experience in terms of immersion and affect seemed to be improved while using the BCI selection method. Furthermore, participants appeared to have more patience when using the BCI selection method than when using the non-BCI selection method, which could have been caused by the curiosity of participants for the BCI selection method or the self overestimation by participants while using the non-BCI selection method.

For future studies it would be interesting to look at long term effects on immersion and affect. Is the BCI selection method still significantly more immersive and affective when participants are used to it? Another question is whether the indulgence towards the BCI selection method is permanent or only temporary. It would also be interesting to add a selection method based on pointing and clicking which was left out of this research. Thereby testing for the cognitive load as the selection methods used in this study required participants to multi-task. Given the intended purpose of this research, it would also be interesting to use commercial BCI input devices (e.g. Emotiv's EPOC). This would also require a proper study comparing the equipment used in this research and commercial BCIs to determine the differences and if they are feasible to use within the setup of this experiment. Other (non-)BCI selection methods, such as speech and gestures could also prove to increase immersion and affect.

Acknowledgments. The authors gratefully acknowledge the support of the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science. The authors would also like to thank Michiel Hakvoort for his technical support on the game and Lynn Packwood for improving the language of this paper.

References

1. Beverina, F., Palmas, G., Silvoni, S., Piccione, F., Giove, S.: User adaptive BCIs: SSVEP and P300 based interfaces. *PsychNology Journal* 1(4), 331–354 (2003)
2. Bin, G., Gao, X., Yan, Z., Hong, B., Gao, S.: ShangkaiGao: An online multi-channel SSVEP-based brain-computer interface using a canonical correlation analysis method. *Journal of Neural Engineering* 6(4), 46002 (2009)
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. Brown, E., Cairns, P.: A grounded investigation of game immersion. In: CHI 2004 Extended Abstracts on Human Factors in Computing Systems, pp. 1297–1300. ACM, New York (2004)
5. Cheng, M., Gao, X., Gao, S., Xu, D.: Design and implementation of a brain-computer interface with high transfer rates. *IEEE Transactions on Biomedical Engineering* 49(10), 1181–1186 (2002)
6. Cooper, R., Osselton, J., Shaw, J.: EEG technology. Butterworths, London (1969)
7. Farwell, L., Donchin, E.: Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology* 70(6), 510–523 (1988)
8. Galán, F., Nuttin, M., Lew, E., Ferrez, P., Vanacker, G., Philips, J., Van Brussel, H., Millán, J.: An asynchronous and non-invasive brain-actuated wheelchair. In: 13th International Symposium on Robotics Research (2007)
9. van Gerven, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., Ramsey, N., Haselager, P., Vuurpijl, L., Gielen, S., Desain, P.: The brain-computer interface cycle. *Journal of Neural Engineering* 6(4), 041001 (2009)
10. Hakvoort, G., Reuderink, B., Obbink, M.: Comparison of PSDA and CCA detection methods in a SSVEP-based BCI-system. Technical Report TR-CTIT-11-03, Centre for Telematics and Information Technology, University of Twente (2011)
11. Hoffmann, U., Vesin, J., Ebrahimi, T., Diserens, K.: An efficient P300-based brain-computer interface for disabled subjects. *Journal of Neuroscience Methods* 167(1), 115–125 (2008)
12. Jennett, C., Cox, A., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., Walton, A.: Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies* 66(9), 641–661 (2008)
13. Lansing, R., Schwartz, E., Lindsley, D.: Reaction time and EEG activation under alerted and nonalerted conditions. *Journal of Experimental Psychology* 58(1), 1–7 (1959)
14. Lin, Z., Zhang, C., Wu, W., Gao, X.: Frequency Recognition Based on Canonical Correlation Analysis for SSVEP-Based BCIs. *IEEE Transactions on Biomedical Engineering* 53(12), 2610–2614 (2006)

15. Lopez, M., Pelayo, F., Madrid, E., Prieto, A.: Statistical characterization of steady-state visual evoked potentials and their use in brain–computer interfaces. *Neural Processing Letters* 29(3), 179–187 (2009)
16. Nijholt, A., Plass-Oude Bos, D., Reuderink, B.: Turning shortcomings into challenges: Brain-computer interfaces for games. *Entertainment Computing* 1(2), 85–94 (2009)
17. Nijholt, A., Tan, D., Allison, B., et al.: Brain-Computer Interfaces for HCI and Games. In: *CHI 2008 Extended Abstracts on Human Factors in Computing Systems*, pp. 3925–3928. ACM, New York (2008)
18. Norman, D.: Emotion & design: attractive things work better. *Interactions* 9(4), 36–42 (2002)
19. Pagulayan, R., Keeker, K., Wixon, D., Romero, R., Fuller, T.: User-centered design in games. In: *The Human-Computer Interaction Handbook*, pp. 883–906 (2002)
20. Picard, R.: *Affective computing*. The MIT press, Cambridge (2000)
21. Reilly, E.L.: EEG Recording and Operation of the Apparatus. In: *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*, pp. 139–160. Lippincott Williams & Wilkins, Baltimore (1999)
22. Reynolds, C.W.: Flocks, herds and schools: A distributed behavioral model. In: *SIGGRAPH 1987 Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 25–34. ACM, New York (1987)
23. Ruen Shan, L., Ibrahim, F., Moghavvemi, M.: Assessment of Steady-State Visual Evoked Potential for Brain Computer Communication. In: *3rd Kuala Lumpur International Conference on Biomedical Engineering 2006*, pp. 352–354. Springer, Heidelberg (2006)
24. Volosyak, I., Cecotti, H., Gräser, A.: Impact of Frequency Selection on LCD Screens for SSVEP Based Brain-Computer Interfaces. In: Cabestany, J., Sandoval, F., Prieto, A., Corchado, J.M. (eds.) *IWANN 2009*. LNCS, vol. 5517, pp. 706–713. Springer, Heidelberg (2009)

Understanding Goal Setting Behavior in the Context of Energy Consumption Reduction

Michelle Scott, Mary Barreto, Filipe Quintal, and Ian Oakley

Madeira Interactive Technologies Institute, University of Madeira, Caminho da Penteada,
Funchal, 9020-105, Madeira, Portugal
{mscott, fquintal}@m-iti.org,
maryluisbarreto@gmail.com, ian@uma.pt

Abstract. Home energy use represents a significant proportion of total consumption. A growing research area is considering how to help everyday users consume less. However, simply *determining* how to best reduce consumption remains a challenging task for many users. Based on goal setting theory, this paper presents two lab studies (based on the presentation of detailed scenarios and the solicitation of goal selections for the individuals depicted) in order to better understand how users make such decisions. It reveals a preference for goals that are perceived to be easy and specific, rather than those known to be effective (e.g. those that reduce energy consumption) or generic. Goal setting theory suggests that easy goals lead to low levels of commitment and motivation, suggesting such choices may be doubly ineffective. Ultimately, this paper contributes to a better understanding of users' goal selections and argues this is a prerequisite to effectively supporting users in reducing resource consumption.

Keywords: Sustainability, Goal-Setting, Motivation, Energy Consumption.

1 Introduction

In the USA, energy consumption in private homes accounts for 22% of total use [14]. With increasing pressure placed on traditional sources and mechanisms of energy generation, there is growing interest in ways to reduce these levels. One way this can be achieved is via the design of interactive systems that encourage, support and motivate individual users to reduce their levels of consumption. Indeed, this is a rapidly developing research area in Human-Computer Interaction (HCI) covering topics as diverse as novel measurement systems [12], the design of sophisticated *eco-feedback* devices [6] and the exploration of how psychologically grounded theories of motivation and behavior change can best be adapted to leverage these rapid technological advances [7].

This paper extends this work. It explores how goal-setting theory, a psychological framework for understanding motivation and behavior change, can be applied to the task of reducing the home energy consumption of everyday users. Although there are numerous previous studies on this topic [e.g. 1, 10], this paper makes two main novel contributions. Firstly, we report data that supports goal setting theory within the

specific domain of sustainability and within an interactive interface. Providing a domain specific validation of this theory will help generate effective new techniques for accurate, real-time capture of consumption activity and the development of advanced systems and interfaces for processing, storing and presenting this material. The second contribution relates to the notion that users have a poor understanding of how to set goals that result in significant changes to consumption levels [2]. We start to explore this idea in detail and show that people tend to pick easier goals for themselves, perhaps because they feel that others will choose difficult goals in their place. This has implications for the design of interactive systems since goal-setting theory posits that difficult goals are more effective motivators than easier goals. In this way, this paper contributes to our understanding of how people make sustainable goal choices, work that has direct application to the domain of HCI and sustainability.

This paper takes steps towards achieving these objectives. Specifically, it describes two studies exploring the nature of the goals users select in home energy reduction scenarios and the feasibility with which they regard them. The method used is a fast, economical and effective way to test theories in this domain and can be applied to other theoretical constructs. By casting light on user's perceptions of appropriate goals in this domain, this paper highlights possibilities for designing systems that encourage and support users in selecting appropriate goals.

2 State of the Art

Goal setting theory is an established and actionable framework for understanding how to motivate behavior change [8]. Fundamentally, it explores how the type and form of goals affect people's level of motivation and ability to achieve targets.

Two of the most important aspects within this framework are the *challenge* and *clarity* of goals. Studies have confirmed that difficult goals promote the highest levels of effort and performance as long as they are clearly expressed and criteria for successful achievement are well identified [3, 8]. Vague goals, such as achieving optimal personal performance in some task, lack external reference and allow for a wide range of performance levels. Specificity decreases performance variability among users by reducing ambiguity with increasingly precise goals.

Other key factors affecting performance towards meeting goals include those that vary among individuals, such as self-efficacy (a measure of perceived empowerment), and goal commitment. People with higher self-efficacy choose more difficult goals for themselves than those with lower self-efficacy. They also have a higher commitment to achieving goals and are better at responding to negative feedback. Feedback is important for goals to be effective; feedback plus goals are more effective than goals alone [8].

Researchers have also applied goal-setting theory to consumption reduction scenarios. Becker [3] tested the effects of combined goal-setting and feedback on conservation behavior. Two groups of 40 households were either given electrical consumption feedback or not, three times a week. The two groups were also split into those with an easy (2%) savings goal or a difficult (20%) savings goal. The results showed that the difficult-goal-plus-feedback group was the only group that used significantly less (13%) electricity.

In contrast, McCalley [10], reported no difference in energy reduction for people who chose 5% and 20% goal levels in a task that gave immediate feedback on energy conserved using a washing machine simulation. Most people saved 20% compared to the control condition; this could suggest that self-set goals are more effective than imposed goals regardless of the goal level. Indeed evidence supports this assertion. For instance, the benefits of setting your own goal are cognitive rather than motivational. Autonomy in goal setting leads to setting higher challenges and having greater performance than when goals are assigned. Once you have chosen your own goal, you are also more committed to that goal [15].

Abrahamse [1] reinforced these points by indentifying the importance of supporting users in selecting personal, specific goals. They looked at integrating goal setting with tailored feedback about energy consumption. Participants received tailored recommendations via a website but their only goal was to reduce 5% of energy consumption. Feedback was given from their self-reports online after 2 and 5 months and households in the study ultimately saved 5.1% energy. This paper expands upon this work by allowing participants to choose their own goals from a set that includes options such as reducing consumption by percentages as well as more concrete goals such as using sleep mode on computers. Using self-chosen and more specific goals may be a more effective technique than setting abstract percentages. Finally, He et al. [7] suggested that adapting goals to specific situations and users is important in creating effective motivational systems.

In summary, this literature suggests that goal-setting theory has much to offer as an actionable framework for designing effective eco-feedback systems that motivate users to reduce consumption. Questions regarding how to select appropriate goals remain, in part due to the diversity of the literature on this topic and in part due to an undeveloped understanding of the basis with which users select goals in this domain [2]. The work in this paper attempts to address these issues via an experimental paradigm that allows users to choose goals with different levels of difficulty and specificity and provides instant feedback on these selections.

3 Sustainable Goals Pilot Study

This study explored how people select goals in order to reduce resource consumption in home scenarios. It builds on prior work suggesting that users typically select inappropriate or ineffective goals [2] and aims to more deeply understand the factors contributing to goal choice.

In order to do this, three scenarios depicting different home settings and lifestyles were developed. Scenarios were chosen because we felt it was more feasible within a lab setting compared to real time feedback. We argue that the use of such scenarios provides a mechanism for standardizing between participants, and is a simple and effective early-stage alternative to real system deployments that gather data about participants' current behaviors and household energy consumption. The scenarios were instantiated as narrated descriptions accompanied by illustrative sketches and produced in a video format. Table 1 highlights key aspects of the scenarios. A set of 11 unique goals was selected for each scenario (33 total). Goals were drawn from a literature review [e.g. 11] and sources such as the StepGreen social network [9].

A representative set of goals is also shown in Table 1. An additional criterion for goal selection was to include goals that varied on level of difficulty (easy/hard) and context (contextual/non-contextual) in each scenario. The easy/difficult categorization was validated using an online questionnaire.

Table 1. Scenario types and corresponding specific goals

	High Income	Low Income	Single Woman
Scenario Details	Doctor and architect, 1 child, 5 bed house, pool	Janitor and supermarket clerk, 3 children, 3 bed house	Lives alone, 1 bed apt, career focused
Specific Goals	Don't heat pool in summer. Turn off lights and take advantage of sunlight.	Repair leaky taps quickly. Don't use standby mode on appliances.	Use public transport to get to work. Use energy saving light bulbs.

The questionnaire listed all 33 goals and asked participants to categorize each on a Likert scale spanning easy to difficult. 20 users completed the questionnaire. Of the 33 goals, two goals initially classified as difficult were perceived to be easy by subjects. After removing these from consideration Cronbach's alpha showed a high level of internal consistency of the remaining goals (difficult goals = 0.72 over 9 goals in total). Similarly, two goals originally classified as easy were rated as difficult by participants. After removal of these, Cronbach's alpha showed the internal consistency of the remaining 20 goals to be high (easy goals = 0.86).

In order to create contextual goals, the scenarios were written to include three contextual hints, which were directly related to three of the eleven goals available to choose for that scenario. For example, in the scenario involving the single woman, the narrated description stated that she used regular (non energy saving) light bulbs. Correspondingly, one of the contextual goals for this scenario was to use energy saving light bulbs. In the high-income family scenario, it was mentioned that they had a pool and one of the contextual goal choices for that scenario was to heat the pool less often to save energy. These goal choices were only available for those scenarios. On the other hand, non-contextual goals for each scenario did not relate to contextual detail provided within the scenarios. Illustrative examples of the goals used in the study can be seen in Table 1.

The main study used the three scenarios and 33 validated goals and was completed by 20 participants recruited via an email advertisement on a popular university forum and the snowball sampling method. Ages ranged from 20 to 34 with an average age of 26.5. 17 of the participants were male and 3 female. All were educated to the graduate level or above; 13 were in full time employment while the remaining 7 were students. Most of the sample was Portuguese (13), 3 were Indian and 1 each was Greek, Swedish, Taiwanese and Venezuelan. 13 were employed, 5 were students and 2 were unemployed. 18 were single and 2 were living with a partner. Household size ranged from 1 to 4 with an average size of 2.65, household income ranged from €10,000 to

€48,000 with an average household income of approximately €27,500. Participants were given €5 compensation for their time.

The experiment started by capturing a baseline measure of environmental concern, a simple questionnaire was developed. The level of overall environmental concern was calculated as the mean of three items measured on five point Likert scales. An analysis of the data captured during the studies conducted in this paper indicated this simple measurement tool exhibited a high level of internal consistency (Cronbach's alpha test reporting 0.79). The three items were as follows:

In your opinion, how serious (severe) is global warming? (Likert scale labels from 'Not serious' to 'Very serious')

I feel my energy consumption is something I... ('Don't need to worry about' to 'Do need to worry about')

I feel worried about the possible effects of global warming. ('Not at all worried' to 'Extremely worried')

Participants then moved to a computer interface (built using Adobe AIR) where they were then exposed to the three scenarios in a fully balanced Latin square design - three participants experienced each of the six possible presentation orders. Directly after watching each video, participants were asked to select goals from the validated list that would best enable the depicted family to reduce their resource consumption. They were provided with immediate feedback on the effectiveness of the goals using a range of typically non-homogeneous metrics (e.g. money saved, or impact on carbon footprint). After selecting four goals per scenario, they were asked to rate whether or not they believed that the goals were realistic by rating whether or not the family described in the scenario would achieve it. They were asked the following for each goal chosen:

"How often do you think the family would commit to each of the following goals you have just chosen?"

The four goals were presented with a Likert scale with the points labeled: Rarely/Never, Occasionally, About half the time, Frequently and Almost Always/Always. Finally, after completing this process for all three scenarios, the experiment closed by asking participants to report how often they engaged in the activities implied by the goals used in the study. This questionnaire included all of the thirty-three goals that had been presented in the interface previously in random order. Participants were asked:

"How often do you perform each of the following energy-saving behaviors when you are in your home? Please select not applicable (N/A) if you do not own an item."

These items were presented with the same Likert scale as previously, including the point "Not applicable". Any "Not applicable" answers were excluded from the analysis. Adding up scores from the final questionnaire created a measure of sustainable lifestyle, higher scores indicating higher current sustainable behaviors. This last measure was intended to separate out what participants felt to be ideal goals, from those that they felt to be realistic goals. This is an interesting comparison to make as people can make different choices for others than they do for themselves. In total, the experiment lasted approximately 25 minutes.

Table 2. Most popular goals chosen for households in the scenarios and selves, means and corresponding standard deviations shown in parentheses. ** Significant at the .01 level * significant at the .05 level.

	Selected for others	Self performs	Mean rating of how often others perform	Mean rating of how often self performs	Feedback over year
Use public transport to get to work	72.2%	33.3%	3.23 (0.83)	1.46 (1.76)**	€ 104 saved
Turn off lights and take advantage of sunlight	72.2%	88.9%	3.62 (1.12)	4.38 (1.12)	€ 38 saved
Turn off lights when not in the room	72.2%	100%	4.54 (0.66)	4.69 (0.48)	€ 6 saved
Turn off water when not using	72.2%	100%	4.31 (0.95)	4.92 (0.28)	€ 145 saved

The study was designed with an exploratory analysis in mind. The overarching goal was to cast light on the types of goals people select, with the expectation that there would be tradeoffs between easy and hard goals and goals that are known to be effective and ineffective. Two formal hypotheses were also generated. The first hypothesis is novel in the sustainability domain and relates to goal context. The second serves to check on internal consistency of the experimental setup and determine whether participants were accurately reporting their attitudes and actions. The hypotheses were:

- H1:** Contextual goals will be chosen more frequently within the consumption reduction scenarios than non-contextual goals.
- H2:** Environmental concern will be positively correlated with self-report of engaging more frequently in sustainable behaviors.

3.1 Results

Table 2 shows the most popular goals chosen for each of the three scenarios, including the information presented to participants about projected savings. Participant's self-report of their own behavior is also shown, as is their assessment of whether or not the individuals depicted in the scenarios would adopt the goals.

The least popular goals chosen for the scenarios, along with corresponding feedback are presented below, with the amount of times each was selected is shown in parentheses:

Use energy saving light bulbs - €27 saved (1)

Eat lunch at home once or twice a week - save on petrol costs and cut emissions (1)

Save 5 Euros from your bill a month - €60 saved (0)

Hypothesis 1 was supported. Goals were coded as either 1 for chosen or 0 for not chosen and means calculated. A t-test ($t(17) = 3.12, p < 0.01$) showed participants selected goals classified as contextual (Mean = 0.44, SD = 0.10) over those goals classified as non-contextual (Mean = 0.32, SD = 0.06). A second t-test ($t(17) = 2.40, p < 0.05$) revealed participants selected easy goals (Mean = 0.41, SD = 0.07) over those rated as difficult (Mean = 0.28, SD = 0.16). The top 4 most popular goals shown in Table 2 were all classified as easy goals.

Hypothesis 2 was also supported. The sum of the participants' ratings of how often they engaged in the goals used in the study was calculated as a measure of the sustainability of their lifestyles. This was strongly correlated with the level of environmental concern (Pearson's $r = 0.53, n = 18, p < 0.05$). The sum of participants' ratings of how often they engaged in the goals that were classed as difficult was also calculated as a measure of their sustainable lifestyles. The level of environmental concern also correlated strongly with the difficult sustainable behaviors that participants reported they performed ($r = 0.62, n = 18, p < 0.01$).

Differences between users recommendations of goals for others and their willingness to adopt them personally are clear in this data. In particular, public transportation was recommended for those depicted in the scenarios much more frequently that it was reported to be personally suitable – a t-test showed this difference to be significant ($t(12) = -3.18, p < 0.01$). In contrast, participants reported themselves more willing to rely on sunlight (as opposed to artificial light) than individuals in the scenarios, this result approached significance ($t(12) = 1.87, p = 0.08$). A similar non-significant trend emerged in ratings for turning off water whilst not in use ($t(12) = 2.13, p = 0.06$). The means and standard deviations for these t-tests are shown in Table 2.

3.2 Discussion

The first finding is the firm support for the hypothesis that people prefer contextual goals to non-contextual ones. Contextual goals are more actionable and are thus more likely to be carried out. The selection of more contextual goals over non-contextual ones shows that people need accurate, relevant and contextualized information when they are choosing goals. Providing information that relates to users specific behaviors and the contexts in which they happen will allow them to select more appropriate goals and ultimately better motivate users towards reducing consumption.

The results also indicate that people are poor at selecting optimal goals in this domain. The second and third most popular goals selected related to home lighting and were easy to accomplish but have little measureable impact. This suggests that people select goals based on the ease with which they can be achieved and seamlessly integrated into their routines and lifestyles. People seem to be more aware of the existence of easier goals and think that they are a fast way to make effective changes and be more sustainable. However, such goals are highly problematic. Not only do they have very limited impact on energy consumption, but the ease with which they can be achieved can lead to reduced levels of motivation [8].

Another finding in this study is that users pick different goals for others than they do for themselves. Furthermore, they overestimate the willingness of others to adopt them.

As shown in Table 2, the public transportation goal was the most popular chosen for the individuals depicted in the scenarios, with participants estimating that this would be performed around half the time. However, when the same people were asked how often they performed this activity, a significantly lower rating was recorded.

Finally, perhaps unsurprisingly, higher levels of environmental concern translated into the adoption of more difficult goals; this could be due to the fact that people with higher levels are aware of the impact of their particular behaviors.

4 Follow Up Goal Setting Study

4.1 Method

A larger second study was conducted in order to build on the findings from the pilot study. It used the same scenarios approach as in the pilot. However, this study did revise several methodological shortcomings present in the pilot.

The number of goals was reduced from thirty-three in the pilot to twelve in the second study. Instead of presenting eleven different goals in each scenario, the same twelve goals were presented in each scenario. This enabled a more direct comparison between scenarios. The scenarios were presented to participants in a random order to control for practice and habituation effects.

The goals were selected in order to cover a broad range of behaviors. From the thirty-three used in the pilot, the twelve in the current study were chosen to include an equal number of easy/difficult goals and of vague/specific goals. Validation of goals in the easy/difficult categorization was completed during the pilot. A similar validation for vague/specific goals was performed as part of this study. This took the form of an online questionnaire in which participants rated the goals on a Likert scale. The scale was scored with the following terms: very vague; somewhat vague; neither vague nor specific; somewhat specific; and very specific. After performing this validation, 12 goals were selected such that three goals fell in to the category pairs of easy/specific, three in easy/general, three in difficult/specific and three in difficult/general. Cronbach's alpha reported the internal consistency of the specific goals as .78 and of vague goals as .70. The final goal list is shown in Table 3.

A significant change from the pilot was the removal of goals that were contextual within the individual scenarios. This addressed one of the methodological issues with the pilot: that contextual goals in each scenario might serve as a confound with the overall, scenario-independent level of specificity of the goals.

33 participants completed the study online. They were recruited via online advertisements for participants on Facebook, via email lists and through an online study website. Participants were not compensated for their time. Ages ranged from 16 to 63 with an average age of 30.5. There were 12 males and 20 females, with one person choosing not to report their gender. Just over half of the sample (17) was educated to degree level or above, 10 people had completed high school and 5 people had completed some college, 1 did not report their education. 16 participants were employed, 13 were students, 3 were unemployed and 1 was retired. Most of the sample were Portuguese (18), 7 were from the U.S.A., 3 were from the U.K. and 1 each were from

Romania and India with the remaining 3 choosing not to report nationality. The majority of the sample was single (22), 6 were married, 2 were living with a partner and 1 was divorced. The household size ranged from 1 to 8 with a mean size of 3 and household income ranged from €12,000 to €250,000 with a mean of approximately €55,000. All participants had Internet access at home.

Table 3. Goal list with levels of difficulty and specificity

Goal	Specificity	Difficulty
Save a percentage of your energy bill over time	Vague	Easy
Reduce carbon footprint	Vague	Easy
Do more outdoor activities	Vague	Easy
Save the environment	Vague	Difficult
Take part in a local environmental organization	Vague	Difficult
Compete with neighbor to be more sustainable	Vague	Difficult
Switch off appliances/lights when not in the room	Specific	Easy
Turn off water when not using it	Specific	Easy
Wash full loads only and where possible at 30 degrees Celsius	Specific	Easy
Use public transport	Specific	Difficult
Replace old large kitchen appliances	Specific	Difficult
Become vegetarian	Specific	Difficult

The experiment began by asking for demographic information including basic details about a participant’s household, such as size and income. A baseline measure of environmental concern was then collected. This was achieved with the New Ecological Paradigm [5], a well-established 15-item measure intended for this purpose. Each item is measured on a 5-point Likert scale and an overall score of environmental concern is derived from the mean of pro-environmental responses.

Participants were then presented with the scenarios used in the pilot study (the high-income family, the low-income family and the single woman) in a random order. The list of 12 goals was presented after each scenario video. When the participant moused over a goal, feedback on the effectiveness popped up for that goal. Feedback was as accurate as possible and the majority was derived from content and tools available on the Stepgreen.org website [9]. Other statistics were taken from trusted sources such as the US Department of Water. Participants were then asked to select three goals from the list of 12 that they felt would enable the family presented in the scenario to reduce their resource consumption. After selecting three goals, they were

taken to the next screen, which showed a summary of the goals chosen with the appropriate feedback for each goal. The participants were then asked:

“How often do you think the family would commit to each of the following goals you have just chosen?”

For each of the three goals chosen, participants had to choose from a 5-item Likert scale with the items: Rarely/Never; Occasionally; About half the time; Frequently and Almost always/Always. This was repeated for three times for each scenario. After the scenarios were completed, participants were again presented with the list of 12 goals and asked:

“How often would you commit to each of the following goals?”

This item was scored exactly as the item above. We asked participants in an open-ended question if they would like to add any more goals/activities that they currently do. The study ended with a well-established 10-item measure of self-efficacy [13].

The experimental hypotheses were as follows:

- H1:** Participants will choose specific/easy goals more often than vague/difficult goals within the scenarios and will also rate themselves as more likely to commit to specific/easy goals compared to vague/difficult goals.
- H2:** Participants will choose easy goals for themselves even though they receive feedback that shows they are ineffective goals.
- H3:** Participants will rate others' commitment to goals as higher than their own.

4.2 Results

The first hypothesis was supported. More people chose specific goals for others within the scenarios than vague goals. This result was significant ($t(32) = 6.87, p < 0.001$); means and standard errors are shown in Figure 1. This supports goal setting theory, which states that people prefer specific goals to vague ones. The results show that more people chose easy goals for others rather than difficult ones. This result was also significant ($t(32) = 5.61, p < 0.001$); means and standard errors are shown in Figure 1.

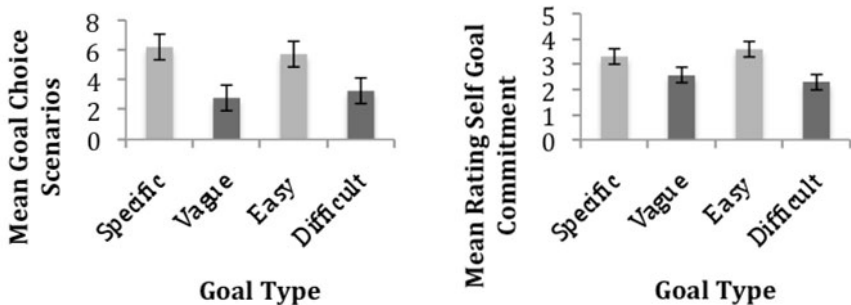


Fig. 1. On the left: Mean goal choices for others in scenarios by goal type (maximum 9). On the right: Mean rating of own commitment by goal type (scale maximum 5). Standard errors are also shown.

Participants also rated themselves as more likely to commit to specific goals than vague goals. This was a significant result ($t(32) = 6.98, p < 0.001$) as shown in Figure 1. There was also a significant difference ($t(32) = 10.37, p < 0.001$) in how often they thought they would commit to easy goals compared to difficult goals (see Figure 1).

Table 4 shows the list of goals in order of popularity over all the three scenarios. The maximum possible for each goal is 99, where each of the 33 participants chose

Table 4. Goal list by scenario popularity. Goal type and feedback presented are also shown.

Goal	Amount chosen	Goal Type	Feedback
Use public transport to get to work	55	Specific Difficult	This could save €104 a year, assuming €10 a week spent on petrol replaced with €2 bus costs a day
Switch off appliances or lights when not in the room	50	Specific Easy	Save €12 yearly and 80 kilos of CO2
Wash full loads at 30C	48	Specific Easy	This could save €13 and 71 kilos of CO2 a year
Save a percentage of energy consumption or money over time	30	Vague Easy	This could save you e.g., 10% from your monthly bill
Turn off water when not using it	27	Specific Easy	This would save €145 yearly and 11,000 liters of water
Do more outdoor activities	25	Vague Easy	Benefit your health, more fresh air and outdoor activities can help prevent diseases and prolong your life
Replace old large kitchen appliances with new energy efficient ones	18	Specific Difficult	The initial investment of new appliance will be recovered within 3 years
Reduce carbon footprint	12	Vague Easy	A collection of several different types of actions, the result would be better for the planet and our natural resources
Compete with neighbor to be more sustainable	12	Vague Difficult	For example, installing a solar panel, this would give you free power for 20 years after the initial cost
Save the environment	9	Vague Difficult	Think about future generations and a better living environment for everyone
Take part in a local environmental organization	7	Vague Difficult	You will get some exercise and fresh air and contribute positively to your community
Be vegetarian	7	Specific Difficult	It's one of the most effective steps you can take and it can save 1600 kilos of CO2 a year

that goal in each of the three scenarios. As can be seen in the table, most of the specific and easy goals are in the top half of the table, as predicted. People prefer goals that are specific, supporting goal setting theory. Feedback for each of the goals is also shown in the table. Two of the most popular goals: switching off appliances and washing full loads of clothes at 30 degrees Celsius, are two of the least effective. These goals save around €1 a month but they were chosen, on average, by half the participants per scenario. This supports our second hypothesis.

There were some small differences in the most popular goals by scenario. In the high-income scenario, the most popular goals were: wash full loads at 30 degrees Celsius (chosen by 21 of 33 participants), switch off appliances (20 out of 33 participants) and use public transport (17 out of 33). In the low-income scenario, the most popular goals were: use public transportation (14 out of 33), turn off appliances, turn off water and do more outdoor activities were all chosen 13 times. In the single woman scenario, the most popular goals were: use public transportation (24 out of 33), wash full loads at 30 degrees Celsius (23 out of 33) and turn off appliances (17 out of 33).

Our third hypothesis was not supported. There were some differences between ratings of how often others would commit to a goal compared to how often the self would commit to a goal. In contrast what was expected and to the findings in the pilot, participants rated themselves as more likely to commit to using public transportation (Mean = 3.18, SD = 1.78) compared to others (Mean = 1.76, SD = 1.03). This was a significant result ($t(16) = -3.67, p < 0.01$). They also rated themselves as more likely to commit to switch off water when it is not being used (Mean = 4.54, SD = 0.78) than others (Mean = 0.85, SD = 1.73). This was a significant difference ($t(12) = -7.22, p < 0.001$). Participants also rated themselves as significantly ($t(16) = 2.28, p < 0.05$) more likely to commit to switching off appliances (Mean = 4.41, SD = 0.51) that are not in use compared to others (Mean = 3.82, SD = 1.01).

Once again there was no significant correlation found between self-efficacy and goal type chosen for others or between self-efficacy and ratings of participant commitment to goal type. Higher self-efficacy was expected to correlate with more difficult goal choices, but this was not found.

There were also no significant correlations found between scores on the New Ecological Paradigm (NEP), a measure of pro-environmental orientation, and types of goal choice for others, or between the NEP and ratings of self-commitment to goals.

4.3 Discussion

Our main study showed full support for our first hypothesis. Participants chose specific and easy goals within the scenarios more often than they chose difficult and vague goals. This confirms the results from the pilot study and also confirms goal setting theory, which states that participants prefer specific goals, as they are more actionable than vague goals. The least popular goal was: be vegetarian. Even though it is a specific goal, this is probably due to that fact that it is too difficult for most people to commit to.

People also indicated they would personally commit more to specific goals rather than vague goals, supporting goal setting theory. The data also showed they would commit more often to easy goals rather than difficult goals, extending the findings

from the pilot. One possible explanation is that difficult goals may be too much of a long-term commitment for people, while easy goals can be rapidly integrated into a person's everyday activities. For instance, both doing more outdoor activities, and taking part in a local environmental organization have clear real world parallels. However, doing more outdoor activities was rated as an easier goal, possibly because it is perceived to be more under a person's direct control than joining an organization. Actionable and effective recommendations are needed for people to choose appropriate sustainable behaviors.

Our second hypothesis was also supported. The two least effective goals in terms of money and CO2 savings were the second and third most popular choices. However, goal setting theory states that difficult and specific goals produce the greatest results. This finding has implications for encouraging behavior change regarding sustainable activities. Easy goals such as turning off appliances or lights tend to be the ones people know most about. People therefore need be provided with more information about effective goals, perhaps presenting more difficult goals or actions in terms of smaller steps that can encourage more effective behavior change over time.

Using public transportation to get to work was the most popular goal and is much more effective than the next two goals in popularity for the scenarios. People need concrete, effective recommendations if they are to make sustainable informed choices about their lifestyle.

There was no support found for our third hypothesis. We expected to see a difference between ratings of commitment to goals between the self and others. There was a difference found but not in the direction expected. More people rated their own commitment to some goals as higher than others' commitment to goals. This is different from the results in the pilot, which suggested that people would choose more difficult goals for others than for themselves. However, the wording of the question was different in the main study: we asked how often participants would be willing to commit to goals, whereas in the pilot we asked about the current behaviors. Our interpretation of this result is that people are more honest when asked about current behaviors and overestimate about their future plans. However, this discrepancy could also be due to differences in the samples; our second sample was broader (and somewhat older) than our first, so they could simply be more aware of the changes they are able to enact in their lives. Further work needs to be done in this area to determine the extent to which people think others will shoulder the responsibility for sustainable energy use.

5 Conclusion

This paper presented two main contributions. Firstly, our studies found support for goal setting theory within the domain of sustainability. Both the pilot and the main study showed that people prefer specific or contextual goals to vague or non-contextual ones. The second contribution shows that users have a poor understanding of how to set goals that have a significant effect on energy consumption levels. Both studies showed that people tend to pick easier goals for themselves, perhaps because they feel that others will choose difficult goals in their place. Since goal setting theory

states that more difficult goals are more effective at getting real results, this has implications for the design of interactive systems.

Previous discussions of goal setting theory have typically been based on aggregate consumption data shown using simple numerical displays [e.g. 10]. While considerable benefits have been shown in this work, this paper argues that additional benefits will emerge through appropriately designed techniques based upon the theory and in the field and would use real-time contextualized feedback [6]. Participants would be able to choose their own goals and receive feedback based on detailed disaggregated data representing consumption practices [4] from their home. They would be presented with customized information and personalized recommendations [7] based on the goals chosen and feedback received. This system would not be annoying, intrusive or repetitive and would adapt to users needs as required. This paper takes steps towards the first design of such a system by showing how users choose goals, the types of goals they choose for themselves and makes an attempt to understand the reason for these choices. This will lead to further work, such as a system based in people's homes, which collects their energy data, allows self-set goals and gives contextualized feedback based on this. The type of goal setting interface utilized in this paper can be used to test theories cheaply and easily. It is quicker than implementing a working system in the field and can be used as a first step to designing useful systems that can have an impact on encouraging sustainable behavior change.

Options for future work on this topic are broad. A key development would be to integrate further work exploring goal selection with real-time sensing and presentation of home energy consumption levels. This will allow the development of interfaces that provide tailored, actionable and contextually relevant goals to users. Regularly updated feedback would also offer users confirmation of the effectiveness of their actions and goals. In summary, this paper has highlighted the need to better understand goal selection behavior in the context of consumption reduction scenarios, so that users can be guided towards more effective and efficient goal selections. Ultimately, this paper suggests that developing a better understanding of users goals will allow the design of better systems to reduce energy consumption.

References

1. Abrahamse, W., Steg, L., Vlek, C., Rothengatter, T.: The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviors, and behavioral antecedents. *Journal of Environmental Psychology* 27, 265–276 (2007)
2. Attari, S.Z., DeKay, M.L., Davidson, C., de Bruin, W.B.: Public perceptions of energy consumption and savings. *Procs. of the National Academy of Sciences of the United States of America* 107, 1–6 (2010)
3. Becker, L.J.: Joint effect of feedback and goal setting on performance: A field study of residential energy conservation. *Journal of Applied Psychology* 63(4), 428–433 (1978)
4. Berges, M., Matthews, H.S., Soibelman, L.: A System for disaggregating Residential Electricity Consumption by Appliance. In: *The IEEE International Symposium on Sustainable Systems and Technology (ISSST)*, Washington, DC (2010)
5. Dunlap, R.E., Van Liere, K.D., Mertig, A.G., Jones, R.E.: Measuring Endorsement of the New Ecological Paradigm: A Revised NEP Scale. *Journal of Social Issues* 56(3), 425–442 (2000)

6. Froehlich, J., Findlater, L., Landay, J.: The Design of Eco-Feedback Technology. In: Proceedings of CHI 2010, Atlanta, Georgia, USA (2010)
7. He, H.A., Greenberg, S., Huang, M.E.: One size does not fit all: applying the Transtheoretical Model to Energy Feedback Technology Design. In: CHI 2010: Sense and Sustainability, pp. 927–936. ACM Press, New York (2010)
8. Locke, E.A., Latham, G.P.: Building a Practically Useful Theory of Goal Setting and Task Motivation. *American Psychologist* 57(9), 705–717 (2001)
9. Mankoff, J., Fussell, S.R., Dillahunt, T., Glaves, R., Grevet, C., Johnson, M., Matthews, D., Matthews, H.S., McGuire, R., Thompson, R.: Stepgreen.org: Increasing Energy Saving Behaviors via Social Networks. In: ICWSM 2010 (2010)
10. McCalley, L.T., Midden, J.H.: Energy conservation through product-integrated feedback: The roles of goal-setting and social orientation. *Journal of Economic Psychology*, 589–603 (2002)
11. Osbaldiston, R., Sheldon, K.M.: Promoting internalized motivation for environmentally responsible behavior: a prospective study of environmental goals. *Journal of Env. Psychology* 23, 349–357 (2003)
12. Patel, S.N., Gupta, S., Reynolds, M.: The Design and Evaluation of an End-User Deployable, Whole House, Contactless Power Consumption Sensor. In: Proceedings of CHI 2010, Atlanta, Georgia, USA (2010)
13. Schwarzer, R., Jerusalem, M.: Generalized Self-Efficacy scale. In: Weinman, J., Wright, S., Johnston, M. (eds.) *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs*, pp. 35–37. NFER-NELSON, Windsor (1995)
14. U.S. Department of Energy. Annual Energy Review 2009. Energy Information Administration, Washington, DC, DOE/EIA-038 (2010)
15. Wright, P.M., Kacmar, K.M.: Goal Specificity as a Determinant of Goal Commitment and Goal Change. *Organizational Behavior and Human Decision Processes* 59(2), 242–260 (1994)

Designing a Context-Aware Architecture for Emotionally Engaging Mobile Storytelling

Fabio Pittarello

Università Ca' Foscari Venezia - DAIS
Via Torino 155,
30172 Mestre (Venezia), Italia
pitt@unive.it

Abstract. This work illustrates the design of a context-aware software architecture supporting the narration of interactive stories for mobile users. The peculiarity of this work is the use of an extended set of context dimensions, including the surrounding environment and the user social network, for enhancing the engagement and the emotional impact on the users experiencing the story.

Keywords: context-awareness, emotionally engaging interaction, mobile devices, social network, storytelling.

1 Introduction

The main goal of this project is the definition of a software architecture for supporting the creation and the delivery of narrations for mobile users, whose evolution is determined by the context. In particular we consider those dimensions of the context - such as the weather, the time of the day or the temperature - that may enhance the emotional impact of the content communicated to the user. Such dimensions often are not considered, because their monitoring may require a set of sensors that usually are not embedded in most devices. The architecture described in this work takes advantage of context data available on the web for overcoming this problem. Besides, the interactive story is here treated as an experience to be virally communicated and shared by a community of users that is considered also as an additional dimension of the context influencing the access to specific fragments of the narration. A small group of potential authors has been involved from the start of the project, before the development of the software prototype. The result was an evolving development process, where the contributions of computer science specialists and users were integrated and produced a result that - in a following preliminary validation phase - was judged satisfying in terms of expressivity, and which is open to further conceptual and technical advances.

2 Related Work

The notion of context has been exploited by several authors [7] [13] that have analyzed and used different dimensions of it, including the location, the user, the device, the network and the time. Context-aware applications guide the user experience in

relation to the context sensed, and they may influence both the interaction and the communication of information. Research on the adaptation of the information presented to the user in relation to the context has been developed particularly for the hypermedia and the web [3] [16], but a wide number of studies have focused also on the domain of the so-called mixed reality, that includes all the different blendings of navigable real and virtual environments. The location has been one of the most explored dimensions of context. Early works include the Active Badge system [22] for locating people inside a building by means of a wearable badge and delivering services to them. Location awareness is used in many applications related to the cultural heritage domain - such as museum guides [11] or educational games for enhancing the visits to archaeological sites [2] - and to tourism. Another dimension of the context, the user history, has been considered for adapting the content presented to the users, for example enabling its proactive presentation in relation to repetitive behavior patterns [6]. Other dimensions, such as the weather or the temperature, have been investigated in a number of research works, including [9] and [15]. As stated in the introduction, in most situations the values related to these variables can't be sensed directly by sensors embedded in the users' devices. That is the reason why some researchers have proposed software architectures [9] that abstract the sensor components and rely on different methodologies for acquiring the values, including the web access. In most cases the knowledge of an extended set of context dimensions is used for informative purposes, such as indicating to tourists specific indoor or outdoor attractions in dependence of bad or good weather conditions. Most of the studies and applications developed so far - especially those ones targeted to mobile devices - have focused on providing appropriate information to the users, lowering their cognitive load, but have been rarely focused on the emotional use of the context. One of the few exceptions is a recent work [20], that emphasizes the emotive potential of the context for augmenting or diminishing the user engagement.

The exam of narratology, that is the study of the structure of the stories, and of its influence on the design of interactive stories has been an important component of our work. An interest survey [5] analyses the theories developed by famous authors from the classical age to contemporaneity - including Aristotle, Propp, Greimas, Barthes and Bremond - in order to find the most suitable for the interactive storytelling domain. Our approach focuses on the studies of an Italian researcher, Cesare Segre [21], that we chose because of its generality and adaptability to different literary genres. Most of the approaches for building models and software architectures for interactive storytelling are focused on the proposal of drama managers (i.e., software components controlling the development of narration on the basis of the story developed by the author) [14] or autonomous agents influencing the evolution of the story [19]. Some of the applications for managing interactive stories have used location awareness for delivering appropriate information to the user. Hansen et al. [12] introduce the concept of location-based Mobile Urban Dramas, where the city becomes the stage for the drama, and the user participates to a play where the actors voices can be heard through her mobile phone headset as she moves through the town. Another interesting project, iLand [8], permits to narrate - to users equipped with mobile devices - stories related to the oral culture and traditions of the island of Madeira, delivering content related to specific locations.

The role of the emotions in the computer human interaction has become increasingly important with the pervasivity of computer systems, that go beyond the limits of work environments. A number of works focus on the so-called affective interaction [17], where the emotional information is communicated by the user to the system in order to improve the interaction. In this work, we share the interest for the user emotions but, rather than studying the means for capturing them, we consider the emotional role that may have the context, associated to a narrative content, for obtaining an emotional engagement of the user.

3 Defining the Conceptual Model

In a previous work [4] the exam of the narrative theories led us to consider the studies of Cesare Segre [21] as one of the most interesting starting points for defining a model of interactive story. The segmentation process of a literary text described by Segre - that leads to identify the main *sequences* of the narration and their relations - was the basis for the model of story defined in [4]. The initial formalization was modified for modeling also stories characterized by a non-linear narrative structure. The Segre's sequence was mapped to the main structure of the interactive story, that we named *scene*. Scenes may originate a complex narrative structure that may be navigated following different paths. The key components of the Segre's sequence were mapped to the two main components of the scene, the *scenery* and the *situation*. A single scene is characterized by a scenery and by one or more situations. The scenery is the passive part of the scene, and it includes the components of the physical environment, such as building and trees. The situation is the active part of the scene, and it corresponds to the Segre's concept of *event*, that represents facts happening inside the story. Each situation is anchored to a specific location inside a scenery and is associated to a definite interactive content. In a certain phase of the narration only a subset of the situations are active and deliver their content when the user enters the associated locations. The entrance of the user and the following interaction usually modify the subset of the active situations, according to the branching structure designed by the story author, and bring the story in a new state. We may classify this model as location-aware and drama manager inspired. In this work we tried to go a step further. Everyone can experience that a fragment of story communicated to the user in the location determined by the author can have a stronger emotional impact.



Fig. 1. Two snapshots of the same location during the day and the night

But the final strength of this impact may greatly be enhanced (or diminished) in relation to the different conditions of time, weather or other dimensions of the surrounding context, as it can be seen in Fig. 1, where the same phrase - *the mad man is wandering through the park* - read in the same location but in different light conditions may cause varying emotional reactions. That is the reason why in this work we enhanced the previous model, increasing the dimensions of the context that regulate the access to the situations. Another feature introduced in this work was the role of the social community in the enjoyment of the narration. While in the previous work we described stories designed for single user interactions, in this paper we considered the possibility for the author to deliver a story for a community of users. The primary role for this community is the sharing of the users' experience, not only for helping other users to discover fragments of content available in specific context conditions, but also for achieving a deeper cognitive and emotional comprehension of the experiences lived by its members. The community represents also an additional dimension of the context that may be used by the story creator for enabling the access to specific fragments of the narration depending on the actions of the community members.

4 Experimenting the Conceptual Model

Because one of the goals of this work was to focus on the narrative expressivity and the emotional engagement of the stories designed with our system, we decided to actively involve from the start a small group of potential authors. We communicated to a small group of students of the Fine Arts Academy of Venice (3 students aged 20) the initial design choices, as described in the previous section, and we asked them to design an interactive story based on an extended set of context types, including location, time, weather and interactions of the users' community. We chose as the location for the story San Servolo, a small island in the Venice lagoon near San Marco square (see Fig. 2 on the left), characterized by a group of buildings immersed in a gracious park. In the past century the island was the site of an asylum for mad men that were treated with various methods, including the cruel electroshock but also the music therapy and the rehabilitative work. The results of the students work came in the form of a story called *San Servolo, travel into the memory of an island*, focused on the life on the island at the times of the sanatorium. The narrative structure proposed by the students was based on a set of ten situations associated to eight locations of the island (see Fig. 2 on the left), often characterized by visible landmarks.

The students designed a set of narrative paths, requiring the access to specific situations before the delivery of the content associated to other situations. Fig. 2 on the right shows a logical scheme displaying, through the use of arrows, where an ordered access to situations is required. Some situations are associated to the extended set of context dimensions described in the previous section: a mad woman of the asylum tells her story next to the sculpture in the park, but only in the afternoons; a piece of classical music - reminder of the music therapy used for the guests of the institution - can be heard by the users facing the south side of the Venice lagoon, but only during the nights characterized by the absence of clouds. Finally, situations 7 and 8 embody the so-called *apparitions*, where the guests of the asylum appear to the users wandering through the park, presenting themselves during the night as mad men and during

the days as men recovered from their illness. The students produced a detailed storyboard for the content of each situation that was the basis for the creation of ten associated video contributions. Their work gave a practical demonstration of the expressive potential of the context-aware story model and indicated where to focus the development. For example the story imagined by the students showed that they were more attracted by the different dimensions of the environmental context rather than by the complexity of the content to deliver. As a matter of fact, the content produced by them was rich from an emotional and narrative point of view, but had a simple technical structure, being composed by single videos for each distinct situation. As a consequence, in the following implementation phase, we focused on the implementation of an expressive set of rules related to the context. We decided to maintain the simple video container suggested by the students for the delivery of the narration. This decision had significant consequences on speeding up the implementation of the user interfaces, helping also their portability towards different client software platforms.

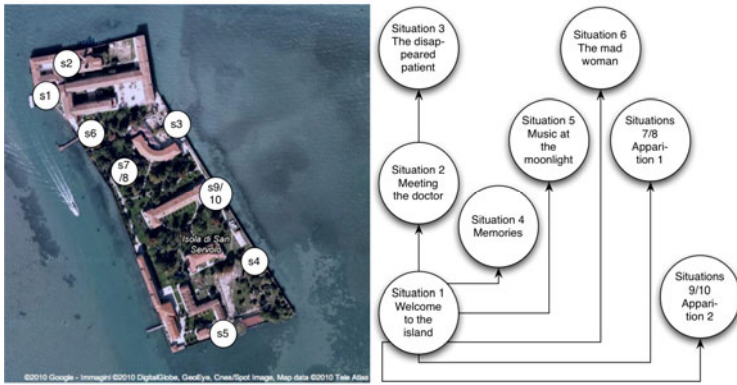


Fig. 2. An aerial view of the island - evidencing the locations associated to the situations defined in the narration - and a logical scheme of the relations between the situations

5 The Software Architecture

Because of the limits of the space we'll focus mainly on the high-level description of the software architecture (Fig. 3). The *director*, the main component of the *story manager* implemented on the server, decides which content to deliver to the client after matching the current values of the context variables with the set of context-related rules defined in the authoring phase for each situation. While the current location value is retrieved from the user device, most of the values of the environmental context are retrieved from different web services. This reasonable solution allows to define a simple and modular software architecture with a noticeable increased number of context dimensions available for narrative purposes. Besides, all the interactions of the user and her community are logged for being matched with the rules related to the user and the community histories. For example the user may access certain situations only if she has already experienced a subset of other situations. Besides, in the current implementation, in order to promote the use of the social network for composing the

different fragments of the narration, the author may specify a rule that inhibits the delivery of content of a certain situation when such content has already been delivered to a given number of users. The discussion space of the community, accessible from the user client interface, takes advantage of the Facebook social plugins [10] for permitting to all the people experiencing the story to write comments and to read the thoughts of the other users. All the comments written by the users authenticated with their Facebook account are visible also in their Facebook profiles, contributing to spread the interest about the interactive story towards their social network (including of course the users that access Facebook from desktop platforms). Generally speaking, in the current implementation the condition for the delivery of the content of a given situation can be informally expressed as follows:

deliver content of situation_i **if** *current location* = location of situation_i
and *current weather* = type of weather_k **and** *min time*_i < *current time* < *max time*_i
and *min temperature*_i < *current temperature* < *max temperature*_i
and (cont. delivery for situation_{i,j1} **and** ... **and** cont. delivery for situation_{i,jn} = true)
and *current num. of deliveries for situation*_i < max number of deliveries for situation_i

where situation_i ∈ set of situations S, type of weather_k ∈ set of weather values W, min time_i, max time_i, min temperature_i and max temperature_i depend from situation_i. The set {cont. delivery for situation_{i,j1}, ..., cont. delivery for situation_{i,jn}} expresses the state of content delivery for a subset of situations R_i ⊂ S and different from situation_i.

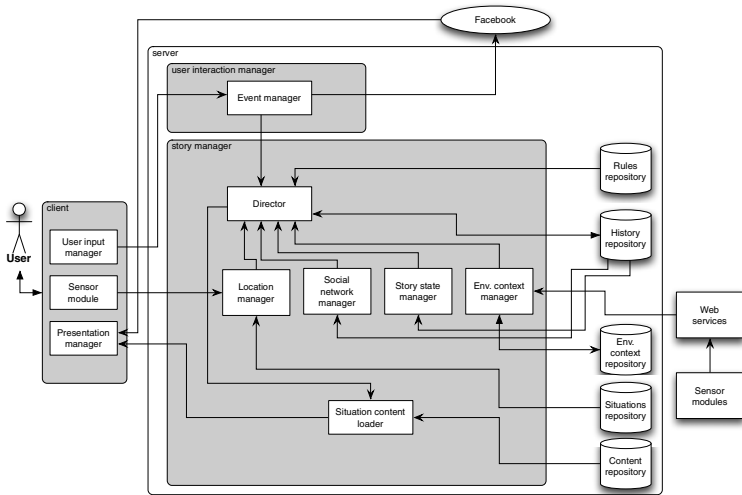


Fig. 3. High-level view of the software architecture

All the components of the proposed architecture have been implemented using standard web technologies, including a JSP application server connected to a PostgreSQL DBMS and to a web server. On the client side, the use of HTML5 has permitted a rapid implementation of the prototypical interface, displayed in Fig. 4.

After the required registration, the user activates the *Discovering San Servolo* function and starts wandering through the San Servolo island for capturing the fragments of the narration. Each time the system recognizes the context conditions for displaying the interactive content associated to a specific situation, such content is delivered to the client, as shown by Fig. 4 on the left. The user can decide at any time to switch to the second function of the client application, *Share your experience* (Fig. 4 on the right), for sharing her thoughts with her social network. The software architecture was preliminary tested with another group of students of the Fine Arts Academy (6 students, aged 20). After the test the users filled in a brief questionnaire. All the users gave a positive evaluation of the narrative expressivity allowed by the software architecture and of the possibility to share the experience. The early involvement of users resulted in a smooth development process, but also stimulated new ideas, as the application of the architecture to new domains. Other groups of users are currently experimenting the application, and we hope to have further interesting feedback for improving the quality of the system. I gratefully acknowledge Alessia Bort, Paola Bressan, Giorgia Franchin and Giorgia Sportelli for their contribution to the elaboration of the case study and to the creation of the software prototype.

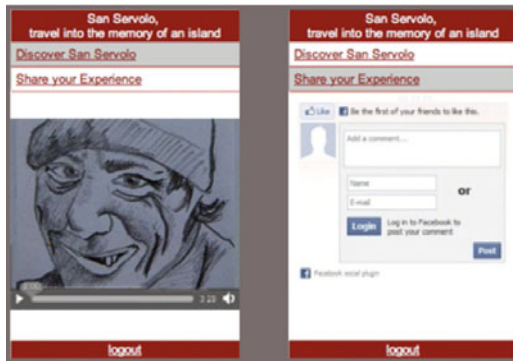


Fig. 4. Screenshots of the browsing interface for mobile devices

References

1. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggle, P.: Towards a Better Understanding of Context and Context-Awareness. In: Gellersen, H.-W. (ed.) HUC 1999. LNCS, vol. 1707, pp. 304–307. Springer, Heidelberg (1999)
2. Ardito, C., Buono, P., Costabile, M.F., Lanzilotti, R., Pederson, T.: Mobile Games to Foster the Learning of History at Archaeological Sites. In: VL/HCC 2007, pp. 81–86 (2007)
3. Brusilovsky, P.: Adaptive hypermedia. *User Modeling and User Adapted Interaction* 11(1-2), 87–110 (2001)
4. Carnielli, A., Pittarello, F.: Interactive Stories on the Net: a Model and an Architecture for X3D worlds. In: Web3D 2009, pp. 91–99. ACM Press, New York (2009)
5. Cavazza, M., Pizzi, D.: Narratology for Interactive Storytelling: A Critical Introduction. In: Göbel, S., Malkewitz, R., Iurgel, I. (eds.) TIDSE 2006. LNCS, vol. 4326, pp. 72–83. Springer, Heidelberg (2006)

6. Celentano, A., Pittarello, F.: Observing and Adapting User Behavior in Navigational 3D Interfaces. In: AVI 2004, pp. 275–282. ACM Press, New York (2004)
7. Chen, G., Kotz, D.: A Survey of Context-Aware Mobile Computing. Technical Report TR2000-381, Dartmouth College, Department of Computer Science (2000)
8. Dionísio, M., Nisi, V., van Leeuwen, J.P.: The iLand of Madeira - Location Aware Multimedia Stories. In: Aylett, R., Lim, M.Y., Louchart, S., Petta, P., Riedl, M. (eds.) ICIDS 2010. LNCS, vol. 6432, pp. 147–152. Springer, Heidelberg (2010)
9. Erdmann, D., Dorfmüller-Ulhaas, K., André, E.: Integrating VR-Authoring and Context Sensing: Towards the Creation of Context-Aware Stories. In: Göbel, S., Malkewitz, R., Iurgel, I. (eds.) TIDSE 2006. LNCS, vol. 4326, pp. 151–162. Springer, Heidelberg (2006)
10. Facebook Social Plugins, <http://developers.facebook.com/plugins>
11. Gordillo, S., Rossi, G., Lyardet, F.: Modeling Physical Hypermedia Applications. In: SAINT-W 2005, pp. 410–413. IEEE Computer Society, Los Alamitos (2005)
12. Hansen, F.A., Kortbek, K.J., Grønbaek, K.: Mobile Urban Drama - Setting the Stage with Location Based Technologies. In: Spierling, U., Szilas, N. (eds.) ICIDS 2008. LNCS, vol. 5334, pp. 20–31. Springer, Heidelberg (2008)
13. Kappel, G., Retschitzegger, W., Kimmerstorfer, E., Pröll, B., Schwinger, W., Hofer, T.: Towards a Generic Customisation Model for Ubiquitous Web Applications. In: Second Int. Workshop on Web Oriented Software Technology, IWOST, pp. 79–104 (2002)
14. Lamstein, A., Mateas, M.: Search-Based Drama Management. In: 2004 AAAI Workshop on Challenges in Game AI. AAAI Press, Menlo Park (2004)
15. Moltchanov, B., Mannweiler, C., Simoes, J.: Context-Awareness Enabling New Business Models in Smart Spaces. In: Balandin, S., Dunaytsev, R., Koucheryavy, Y. (eds.) ruSMART 2010. LNCS, vol. 6294, pp. 13–25. Springer, Heidelberg (2010)
16. Perkowitz, M., Etzioni, O.: Adaptive Web Sites. *Communications of the ACM* 43(8), 152–158 (2000)
17. Picard, R.W.: *Affective Computing*. MIT Press, Cambridge (1997)
18. Riedl, M.O., Stern, A.: Failing Believably: Toward Drama Management with Autonomous Actors in Interactive Narratives. In: Göbel, S., Malkewitz, R., Iurgel, I. (eds.) TIDSE 2006. LNCS, vol. 4326, pp. 195–206. Springer, Heidelberg (2006)
19. Sanchez, S., Balet, O., Luga, H., Duthen, Y.: Autonomous Virtual Actors. In: Göbel, S., Spierling, U., Hoffmann, A., Iurgel, I., Schneider, O., Dechau, J., Feix, A. (eds.) TIDSE 2004. LNCS, vol. 3105, pp. 68–78. Springer, Heidelberg (2004)
20. Schöning, J., Bartindale, T., Olivier, P., Jackson, D., Krüger, A., Kitson, J.: iBookmark: Locative Texts and Place-based Authoring. In: CHI 2009 - Spotlight on Works in Progress, pp. 3775–3780. ACM Press, New York (2009)
21. Segre, C.: *Introduction to the Analysis of the Literary Text*. Indiana University Press (1988)
22. Want, R., Hopper, A., Falcao, V., Gibbons, J.: The Active Badge Location System. *ACM Transactions on Information Systems* 10(1), 91–102 (1992)

Towards Emotional Interaction: Using Movies to Automatically Learn Users' Emotional States

Eva Oliveira^{1,2}, Mitchel Benovoy³, Nuno Ribeiro⁴, and Teresa Chambe²

¹ Digital Games Research Centre, Polytechnic Institute of Cávado and Ave,
4750-810 Barcelos, Portugal
eoliveira@ipca.pt

² LaSIGE, University of Lisbon FCUL, 1749-016 Lisbon, Portugal
tc@di.fc.ul.pt

³ Center for Intelligent Machines, McGill University, Montreal, Quebec, Canada
benovoy@cim.mcgill.ca

⁴ CEREM – Centro de Estudos e Recursos Multimediáticos, Universidade Fernando Pessoa
nribeiro@ufp.edu.pt

Abstract. The HCI community is actively seeking novel methodologies to gain insight into the user's experience during interaction with both the application and the content. We propose an emotional recognition engine capable of automatically recognizing a set of human emotional states using psychophysiological measures of the autonomous nervous system, including galvanic skin response, respiration, and heart rate. A novel pattern recognition system, based on discriminant analysis and support vector machine classifiers is trained using movies' scenes selected to induce emotions ranging from the positive to the negative valence dimension, including happiness, anger, disgust, sadness, and fear. In this paper we introduce an emotion recognition system and evaluate its accuracy by presenting the results of an experiment conducted with three physiologic sensors.

Keywords: Affective computing, Emotion-aware systems, Human-centered design, Psychophysiological measures, Pattern-recognition, Discriminant analysis, Support vector machine classifiers, Movies classification and recommendation.

1 Introduction

Society's relation with technology is changing in such ways that it is predictable that, in the next years, Human Computer Interaction (HCI) will be dealing with users and computers that can be anywhere, and at anytime, and this changes interaction perspectives for the future. Human body changes, expressions or emotions would constitute factors that became naturally included in the design of human computer interactions [1]. HCI aims to understand the way users experience interactions and strives to stimulate the sense of pleasure and satisfaction [2] by developing systems that focus on new intelligent ways to react to user's' emotions. In fact, our cognition, creativity, health and aesthetic sensibility are highly influenced by our emotions,

playing an important role in our lives. Human Computer Interaction research community has been using physiologic, brain and behavior measures to study possible ways to identify and use emotions in human-machine interactions [3; 4]. However, there are still challenges in the recognition processes, regarding the effectiveness of the mechanisms used to induce emotions. The induction is the process through which people are guided to feel one or more specific emotions, which provokes body reactions. If the induction is not well succeed it would be more difficult to detect or recognize that changes. In this work, we present a novel pattern recognition system, based on discriminant analysis and support vector machine classifiers, which is validated using movies' scenes selected to induce emotions ranging from the positive to the negative valence dimension, including happiness, anger, disgust, sadness, and fear. The recognition engine was designed as a low-cost emotions recognition system and it is part of a system - iFelt - which is an interactive web video system developed to learn user's emotional patterns using movies' scenes selected to induce emotions. In this paper we present our recognition method and the accuracy results by presenting an experiment with three physiologic sensors. Following this introduction, section 2 makes a review of most relevant related work, and section 3 introduces the iFelt system. Section 4 introduces the iFelt experiment, with a focus on the elicitation method. The paper ends in section 5, with conclusions and a discussion of perspectives for future work.

2 Related Work

There are different approaches for the affective evaluation of interactions because emotions are not straightforward to understand and are very complex to model. There are evaluations based on emotional dimensions, such as valence [15] or based in categories [16]. The recognition of emotions through physiological patterns is an unconscious way to assess emotions, which can be even more accurate than self-assessment [5]. However, recognition processes commonly categorize emotions. Thus, the mapping of physiological patterns into emotional labels relies on the application of one or more emotional models. Based on directives from the Humaine Network of Excellence – an European Union initiative devoted to the evaluation of affective systems - there are three commonly used such models: 1) the Categorical model, which defines emotions as discrete states, 2) the Dimensional model, which bases emotion perspectives in a spatial circumplex of emotion properties (the most common refer to arousal and valence), and 3) the Appraisal model, which defends that emotions depend on people's own evaluation of events and its circumstances. The characterization of emotions by physiologic patterns faces some problems, but it has some advantages when compared to other recognition methods.

The characterization of emotions has still some limitations regarding the differentiation of emotions from physiologic signals, namely, in finding the adequate elicitation to target a specific emotion [4]. Moreover, due to the fact that being emotions, time-, space-, context- and individual-based, trying to find a general pattern for emotions, and trying to obtain a "ground truth" can be difficult: these are some of the major problems we currently face [3]. On the other hand, there is a major advantage when compared to facial and vocal recognition, because emotions cannot be intentional

- we cannot trigger the autonomic nervous system (ANS) contrarily to the so called “poker face” where people disguise facial expressions as well as vocal utterances [13]. Another common argument against physiologic measurements for emotion analysis is the fact that sensors are invasive, but new wearable sensors and related technological advances promise to allow for mouses that can incorporate sensors in the same way as, for example, the ones used in wrist band’s to measure affective states [6]. Regarding the matter investigated in this work, Picard et al. also claim that cameras for facial recognition can be more invasive than physiologic sensors because they reveal identity or appearance besides emotional information. Films are by excellence the form of art that exploits our affective, perceptual and intellectual activity. In 1996, a research group [8] tested eleven induction methods and concluded that films are the best method to elicit emotions (positive and negative) and mainly when subjects (not studying psychology) are treated individually and are introduced to the purpose of the study. More recently, other researchers used films to induce emotions with different goals. One of the first works is from [8], which tried to find films that induce differential emotional states (dimensional space) while J. Gross et al. [5] tried to find as many films as possible to elicit discrete emotions and find the best films for each discrete emotion. In light of the evidence of distinct physiological responses of emotion, the machine learning and HCI communities have each investigated the automatic recognition of emotions. Picard et al. [6] pioneered this area by showing that some emotional states can be recognized automatically using physiological signals and pattern recognition methods. In [18], the authors made an overview represented in a table (Table 1) that we complete with the elicitation method used, which presents the most relevant studies regarding emotion recognition, using different physiologic signals and different classification methods. Methods that have used movies had the lower results, which in our opinion demands for new ones that can more effectively explore the power of movies to induce emotions given that, according to [8], movie scenes are considered to be one of the best methods to induce emotions.

Table 1. Four studies on automatic physiological-driven classification on affect

Ref	Year	Signals	Participan Features		Select./Red.	Classifiers	Target	Results	Elicitation
[6]	2001	C,E,R,M	1	40	SFS, Fisher	LDA	8 emotions	81%	images
[3]	2008	C,E,R,M	3	110	SBS	LDA	4 emotions	70%	music
[14]	2008	C,E,R,M	40	5	-	SVM	5 emotions	47%	movies
[7]	2009	C,E,R,EE	10	18	-	LDA, SVM,RVM	3 emotions	51%	movies

Signals: C: cardiovascular activity; E: electrodermal activity; R: respiration; M: electromyogram and;EE: electroencephalographic; Selection: SFS: Sequential Forward Selection; SBS: Sequential Backward Selection; Fisher: Fisher projection; Classifiers: SVM: Support Vector Machine; RVM: Relevance Vector Machines; LDA: Linear Discriminant Analysis;

The collection of physiologic data when users are watching movies was recently developed in the psychology area to test whether films can be efficient emotional inductors, which could help psychologists in specific treatments [10], or in the computer science area to automatically summarize videos according to the emotional impact on their viewers [11,12]. Money & Agius (2009) [12] report an experiment on

how user physiological responses vary when elicited by different genres of video content in order to validate the development of personalized video summaries. They also showed specific video segments to a group of viewers monitored with biometric artifacts (electro dermal response, respiration amplitude, respiration rate, blood volume pulse and heart rate) and concluded that there are significant differences between users when watching the same video segments. Money et al. (2009) [12] make use of the dimensional theory to classify the affective results in this study.

In the next section we describe the classification procedure of our system.

3 The iFelt Classification and Recognition Engines

iFelt is an interactive web video system designed to learn users emotional patterns, and explore this information to create emotion based interactions. The system is composed of two components. The “Emotional Recognition and Classification” component which performs emotional recognition and classification of user’s emotional states and the “Emotional Movie Access and Exploration” component that explores ways to access and visualize videos based on their emotional properties and users’ emotions and profiles. In this paper we are focused on evaluation of the Emotional Recognition and Classification component whereas the Emotional Movie Access and Exploration are thoroughly described in [17]. In this section we will describe in detail the elicitation procedure and how we collect, process, and classify biosignals to learn and later recognize emotional patterns as a low-cost emotions recognition system.

3.1 Emotional Elicitation

Every emotion recognition process needs to address the problem of how to induct emotions. Our emotional recognition and classification component is grounded in the induction of emotional states by having users watch movie scenes. The selection of the movie scenes was based in our own judgment and also based on J. Gross’ work [5]. We have selected 3 scenes from their work, and a selection of 13 additional scenes that in our opinion were emotionally intense and represent the set of emotions needed to test our recognition engine, with an average duration of 2 minutes 22 seconds per scene. The system, iFelt, uses the subjects’ data obtained while watching movie scenes to create an engine to enhance automatic recognition of users’ emotional states. The selected movie scenes induct subjects to feel five basic emotions (happiness, sadness, anger, fear and disgust) and the neutral one, so, every subject should see 16 scenes (four of happiness, four of sadness, four of fear, two of disgust and two of anger) and one neutral scene. Based on their feedback, we associated the captured physiological signals with emotional labels, and trained our engine.

3.2 Biosignal Capture

Biosignal recording uses biosensors for measuring Galvanic Skin Response (GSR), Respiration (Resp) and Electrocardiogram (ECG) and is responsible for users’ biosignals recording and signal processing pipeline each sampled at 256 Hz. These

sensors were specifically chosen as they record the physiological responses of emotion, as controlled by the autonomous nervous system, and also because they were already proven sufficient for measuring our five basic emotions (happiness, sadness, anger, fear and disgust) [4,12]. From the heart we measured heart rate, heart rate acceleration, and heart rate variability. We also measured the first and second derivative of the heart rate. For respiration, we measured the rate, the amplitude and the first and second derivative of the rate. For GSR, we measured the stats plus the number of Skin Conductance Responses (SCRs).

3.3 Emotional Classification

Emotionally relevant segments of the recordings that are free of motion artifacts are hand-selected and labeled with the help of the video recordings and subjects responses. High-frequency components of the signals are considered to be noise and filtered with a Hanning window [13]. For the GSR signal, a cutoff frequency of 2.0 Hz was used, whereas for the ECG and respiration signals, cutoffs of 128 Hz and 10 Hz were used in the filters. To account for inherent physiological differences between participants, the mean of the 3-minute silent baseline data preceding each stimulus onset was subtracted from the *active* data and the signal range was adjusted to a [0;1] interval. We extract six common statistical features from each type of the filtered biosignals, of size N ($X_n, n \in [1...N]$): the filtered signal mean, the standard deviation of the filtered signals, the mean of the absolute values of the first differences of the filtered signals, the mean of the absolute values of the first differences of the normalized signals, the mean of the absolute values of the second differences of the filtered signals and the mean of the absolute values of the second differences of the normalized signals. A total of $6 \times 3 = 18$ features are computed from the three types of biosignals. These features were chosen to cover the typically measured statistics in physiological recordings. The advantage of using relatively simple statistical feature is that these can be computed efficiently, opening the door for real-time applications. We employed digital signal processing and pattern recognition, inspired by statistical techniques used by Picard [6], in particular in our use of *sequential forward selection* (a variant of sequential floating forward selection). We specifically chose statistical features, as these are computationally easy to produce, which opens the way to future real-time systems, choosing only classifier-optimal features, followed by *Fisher dimensionality reduction*. For the classification engine, however, we implemented linear discriminant analysis (LDA) rather than the maximum *a posteriori* used by Picard, since our previous experiments with physiological data has shown that LDA produces high classification rates. LDA was selected to dimensionally reduced data by building a statistical model for each emotional class and then cataloguing novel data to the model that best fits. We are thus concerned with finding which classification rule (discriminant function) best separates the emotion classes. LDA finds a linear transformation Φ of the x and y axes that yields a new set of values providing an accurate discrimination between the classes. The transformation thus seeks to rotate the axes with parameter ν , so that when the data is projected on the new axes, the difference between classes is maximized.

3.4 Emotion Recognition

The pattern recognition module uses discriminant analysis, support vector machine (SVM) and k-Nearest Neighbour (K-NN) classifiers [20] to analyze the physiological data and it was validated by the usage of specific movie scenes selected to induce particular emotions. We used the greedy sequential forward floating selection (SFFS) algorithm to form automatically a subset of the best n features from the original large set of m ($n < m$). SFFS starts with an empty feature subset and, on each iteration, exactly one feature is added. To determine which feature to insert, the algorithm tentatively adds to the candidate feature subset one that is not already selected and tests the accuracy of a k -NN classifier built on this provisional subset. A feature that results in the highest classification accuracy is permanently included in the subset, while a poor feature is deleted. The process stops after an iteration where no feature additions or deletions cause an improvement in accuracy. The resulting feature set is now considered optimal. The SVM classifier generates parallel separating hyperplanes that maximize the margins between the subjects' data, which has the effect of minimizing generalization error. Because SVMs are binary classifiers by nature, we used a one-versus-all decision strategy to perform multiclass classification. This technique divides the single multiclass problem into c binary classifiers in which the one with the highest recognition confidence value assigns the final label.

We trained the SVMs with a Radial Basis Functions (RBF) kernel for which the parameters were determined concurrently using an iterative grid selection technique that finds the best combination using the training error of the classifier as a performance metric [20]. The SVM module outputs the identity of the recognized person, along with a classification confidence value based on the distance between the feature vector of the probe and the hyper-margin of the closest subject. The k -NN classifier used here classifies a novel object r by a majority of "votes" of its neighbors, assigning to r the most common class amongst its k nearest neighbors, using the Euclidean distance as metric. It was found through experimentation that a value of $k = 5$ resulted in the best possible selected feature subset.

4 Evaluation and Discussion

We used a portable system - Nexus 4 - with 3 inputs channels for ECG, Respiration and HR. It is a wireless 'real-time' data link computer, and can store up to 24 hours of physiological data on its built-in flash memory. The computer software used to process data was Biotrace¹. Eight participants, averaged 34 year, were submitted to an experiment. Because we presented full length feature films to our subjects, we were able to recruit only eight subjects. However, we deem this number of participants acceptable as a first attempt to classify their emotional reactions. After the subject arrived, the electrodes were attached and the recording system checked. We developed a web interface to easily perform the learning procedure, which began by asking the user to rest for 3 minutes, in a quiet mode. At the beginning of the sessions, a 3-minute silent baseline was recorded, while the participants engaged in focused relaxation by limiting their concentration to their respiration. This pre-stimulus

¹ <http://www.mindmedia.nl/english/biotrace.php>

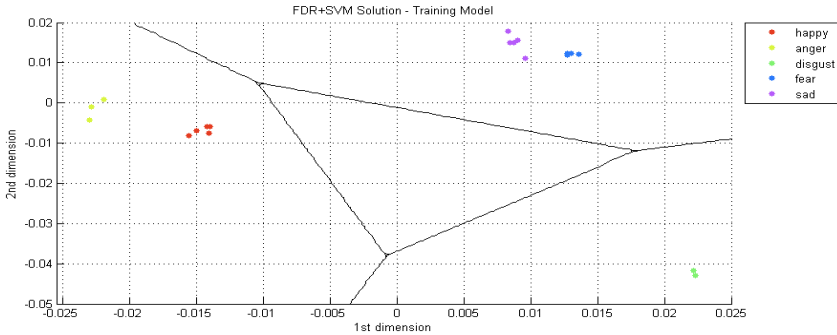


Fig. 1. Class Clustering of Five Emotional States

relaxation time was critical to stabilize the physiology to a homeostatic state, as some participants initially exhibited anxiety at being the focus of attention and being *wired* to the sensors. Then the neutral scene was shown to collect the neutral state right after the baseline. The sequence of 16 movie scenes began with the happiest scenes first alternating with fear related ones. Then, we showed the disgust related scenes alternating with fear related scenes. The last set included the sad scenes, alternating with anger, but chosen in a way that the most intense sad emotions were presented towards the end of the experimental session following recommendations of [5]. We also chose this sequence because it was evident from our previous sessions that when people watched intense sad emotions they became emotionally unavailable to be amused, or even frightened, by any movie scene. In turn, this choice was found to be the most adequate sequence in order to improve recognition accuracy. Upon watching completion, subjects answered a questionnaire describing the emotion from the set which they considered was the dominant, in which intensity and if they enjoyed watching it. After the movies session, the experimenter labeled minute per minute every movie used in this experience with the expected emotion, to compare with the results of the recognition. The learning phase expressed in Figure 1 demonstrates the class clustering of five emotional states: happiness, anger, sadness, fear and disgust projected on the 2D Fisher space along with the SVM class-boundaries of one person, which is satisfactory once we want to learn each user's emotional patterns. It is a positive result because it shows that emotion instances distribution of the same category are near to each other, and apart from the other, which reveals coherence and confidence. To evaluate the pattern recognition process, we compared the expert-labeled movies with the output of the automatically classified data from the iFelt system. The expert labeled the movies in consecutive blocks of one minute length. To compare with the output of the recognition system, which produced classification scores on consecutive five-second windows, the median score was computed over minute long segments of the classified data. At least two subjects watched each of the eight movies, and were classified by the system. With the SVM classifier, the overall average recognition rate is 69% (s.d. 5.0%), which represents a 49% improvement over random choice. Because we are classifying 5 classes, the random probability is 1/5 (20%), which is what the simplest classifier could do. However, our classifier performs at 69%, which is 49% higher than the random

choice of 20%. Our k-NN classifier produced an overall average recognition rate of 47% (s.d. 9.3%). The SVM classification score shows promise that the iFelt recognition system can be used to automatically evaluate human emotions. Although it is important to note that further research needs to be conducted to optimize both the classification algorithms for the type of data used here and the scenes used to learn the emotional patterns. However, two key positive aspects of the system emerged which is the use of easily computed statistical features, which can be used to develop real-time classification systems, and a quite reasonable recognition rate, with only three sensors when compared with the works listed in section 2.

5 Conclusion and Future Work

We presented an emotional recognition system capable of automatically recognizing a set of human emotional states using psychophysiological measures and pattern recognition techniques based on discriminant analysis and support vector machine classifiers. Regarding our experiment methodology we concluded that, every time a subject watched an intense sad movie scene, such as atrocities performed in genocides, the person could not feel happiness in any kind of subsequent happy scene. Another conclusion we made was that people, after watching a very disgusting scene, such as watching sputum in a very expository way, normally kind of forgot the past sensations (sadness, happiness, fear) and felt a little indisposed. The third observation we made was that after 16 movie scenes, in a total duration of 40 minutes, the experience became emotionally intense and subjects became tired and lost their good mood. In order to test the performance of our system, a novel emotion elicitation scheme, based on emotions induced by watching selected movie scenes was presented, engendering a moderate degree of confidence in collected, emotionally relevant, biosignals. Discrete state recognition via physiological signal analysis, using pattern recognition and signal processing, was shown to be reasonably accurate. A correct average recognition rate of 69% was achieved using sequential forward selection and Fisher dimensionality reduction, coupled with a Linear Discriminant Analysis classifier. An important conclusion of this work is that it is easily computed statistical features can be implemented in real-time classification systems, which allows moving towards an emotional interaction system, and also reveals that few features can achieve pretty good results. Even though physiologic sensors are invasive, recent technological advances is resulting in the development of wearable sensors less intrusive, which make our recognition engine useful in assessing human emotional states during human-computer interactions and further validates the use of movies as powerful emotional triggers. Our ongoing research also intends to support *real-time* classification of discrete emotional states, adding also arousal/valence mappings from biosignals for multimedia content classification and user interaction mechanisms by developing emotional aware applications that react in accordance to user's' emotions. In the context of our work we are considering using emotion recognition to automatically create emotional scenes, recommend movies based on the emotional state of the user and adjust interfaces according to user's emotions and based on emotional regulation theories. By creating emotional profiles for both movies and users, we are developing new ways of discovery interesting emotional

information in unknown or unseen movies, compare reactions to the same movies among other users, compare directors intentions with users effective impact, analyze over time our reactions or directors tendencies. Measuring physiological signals of users is a natural input mechanism that can be used to automatically classify information with emotional semantic which can enhance retrieval systems by adding emotional information to search engines, as can be used in interactive applications that take into account the affective states of users.

Acknowledgements. This work is partially supported by FCT through LASIGE Multiannual Funding, PROTEC 2009 (SFRH/BD/49475/2009) and VIRUS research project (PTDC/EIA-EIA/101012/2008).

References

1. Being Human: Human-Computer Interaction in the Year 2020 (2007), <http://research.microsoft.com/hci2020/>
2. Dix, A., Finlay, J.E., Abowd, G.D., Beale, R.: Human-Computer Interaction, 3rd edn. Prentice Hall, Englewood Cliffs (December 2003)
3. Kim, J., André, E.: Emotion Recognition Based on Physiological Changes in Listening Music. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 30(12), 2067–2083 (2008)
4. Maaoui, C., Pruski, A., Abdat, F.: Emotion recognition for human-machine communication. In: 2008 IEEEERSJ International Conference on Intelligent Robots and Systems, pp. 1210–1215 (2008)
5. Rottenberg, J., Ray, R., Gross, J.: Emotion elicitation using films. In: *The Handbook of Emotion Elicitation and Assessment* (2007)
6. Picard, R.W., Vyzas, E., Healey, J.: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(10), 1175–1191 (2001g)
7. Chanel, G., Kierkels, J.J.M., Soleymani, M., Pun, T.: Short-term emotion assessment in a recall paradigm. *Int. J. Hum.-Comput. Stud.* 67, 8 (2009)
8. Westermann, R., Spies, K., Stahl, G., Hesse, F.W.: Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology* 26(4), 557–580 (1996j)
9. Philippot, P.: Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cognition & Emotion* 7(2), 171–193 (1993h); Gross, J.J., Levenson, R. W.: Emotion elicitation using films. *Cognition & Emotion* 9(1), 87–108 (1995i)
10. Kreibig, S.D., Wilhelm, F.H., Roth, W.T., Gross, J.J.: Cardiovascular, electrodermal, and respiratory response patterns to fear- and sadness-inducing films. *Psychophysiology* 44(5), 787–806 (2007)
11. Soleymani, M.S., Chanel, C.G., Kierkels, J.K., Pun, T.P.: Affective Characterization of Movie Scenes Based on Content Analysis and Physiological Changes. In: *International Symposium on Multimedia*, pp. 228–235 (2008)
12. Money, A.G., Agius, H.: Analysing user physiological responses for affective video summarization. *Displays* 30(2), 59–70 (2009)
13. Cliffs: *Discrete-Time signal processing*. Prentice-Hall, New Jersey (1989)
14. Lichtenstein, A., Oehme, A., Kupschick, S., Jürgensohn, T.: Comparing Two Emotion Models for Deriving Affective States from Physiological Data. In: Peter, C., Beale, R. (eds.) *Affect and Emotion in Human-Computer Interaction*. LNCS, vol. 4868, pp. 35–50. Springer, Heidelberg (2008)

15. McQuiggan, S., Lee, S., Lester, J.: Predicting user physiological response for interactive environments: An inductive approach. In: Proceedings of the 2nd Artificial Intelligence for Interactive Digital Entertainment Conference, pp. 60–65 (2006)
16. Isbister, K., Höök, K., Sharp, D., Laakolahti, J.: The Sensual Evaluation Instrument: Developing an affective evaluation tool. In: Proceedings of ACM CHI (Conference on Human Factors in Computing), Montréal, Québec, Canada (2006)
17. Oliveira, E., Martins, P., Chambel, T.: iFelt: Accessing Movies Through Our Emotions. In: EuroITV 2010, 9th European Conference on Interactive TV and Video, ACM SIGWEB, SIGMM & SIGCHI, Lisboa, Portugal, June 29-July 01, 10pgs (2011)
18. Van der Zwaag, M.D., van den Broek, E.L., Janssen, J.H.: Guidelines for biosignal driven HCI. In: ACM CHI2010 Workshop - Brain, Body, and Bytes: Physiological User Interaction, Atlanta, GA, USA (April 11, 2010)
19. Bishop, C.M.: Pattern Recognition and Machine Learning, 740 pages. Springer, Heidelberg (2006)

Motion and Attention in a Kinetic Videoconferencing Proxy

David Sirkin², Gina Venolia¹, John Tang¹, George Robertson¹, Taemie Kim³,
Kori Inkpen¹, Mara Sedlins¹, Bongshin Lee¹, and Mike Sinclair¹

¹Microsoft Research

²Stanford University

³MIT

sirkin@cdr.stanford.edu,
{ginav,johntang,ggr,kori,a-marase,
bongshin,sinclair}@microsoft.com,
taemie@media.mit.edu

Abstract. Compared to collocated interaction, videoconferencing disrupts the ability to use gaze and gestures to mediate interaction, direct reactions to specific people, and provide a sense of presence for the satellite (i.e., remote) participant. We developed a kinetic videoconferencing proxy with a swiveling display screen to indicate which direction that the satellite participant was looking. Our goal was to compare two alternative motion control conditions, in which the satellite participant directed the display screen's motion either explicitly (aiming the direction of the display with a mouse) or implicitly (with the screen following the satellite participant's head turns). We then explored the effectiveness of this prototype compared to a typical stationary video display in a lab study. We found that both motion conditions resulted in communication patterns that indicate higher engagement in conversation, more accurate responses to the satellite participant's deictic questions (i.e., "What do *you* think?"), and higher user rankings. We also discovered tradeoffs in attention and clarity between explicit versus implicit control, a tension in how motion toward one person can exclude other people, and ways that swiveling motion provides attention awareness, even without direct eye contact.

Keywords: Video-mediated communication, videoconferencing, gaze awareness, proxy, telepresence.

1 Introduction

Attention is fundamental to the flow of face-to-face conversations. Each participant projects cues of what he is paying attention to and other participants interpret these cues to maintain awareness of his locus of attention. This awareness helps them understand his deictic references. Both production and consumption of awareness cues occur at conscious and subconscious levels [1].

Videoconferencing systems disrupt the link between attention projection and attention awareness. They do this in part because they do not faithfully reproduce the

spatial characteristics of gaze, body orientation, and pointing gestures. This disruption is one of the reasons why video-mediated communication is less effective than face-to-face interaction. The lack of a shared physical environment further hampers participants' abilities to use spatial cues to support conversation and direct attention [2, 3]. Videoconferencing configurations that involve multiple people at one site offer multiple plausible loci of attention, increasing the potential for confusion.

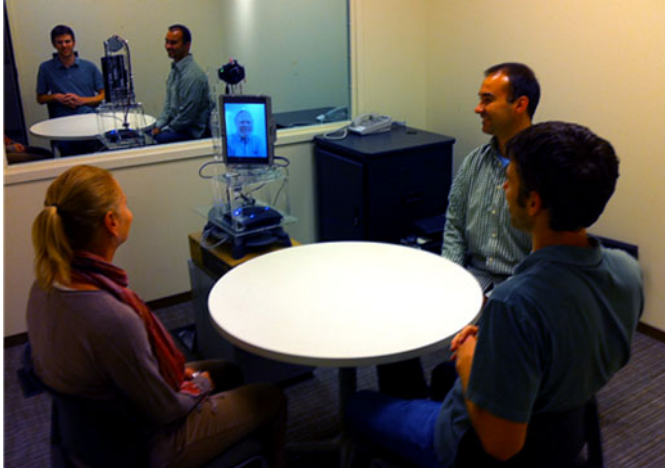


Fig. 1. The experimental setup, showing three collocated participants and the kinetic proxy seated around a table. The proxy was operated by a confederate, and positioned so that its swiveling display approximately matched participant eye-height.

We are particularly interested in videoconferencing systems to support *hub-and-satellite* meetings, where most participants are collocated except for one participant at a satellite location. This satellite is represented in the collocated space by a *proxy* device consisting of a display screen, camera, speaker, and microphone (Fig. 1). The satellite perceives the hub location through streams of audio and video displayed on his computer display.

A study of proxies in everyday use has documented the benefits of a physical representation of the satellite in group interaction [4]. Our own use and studies of proxies in our day-to-day work has led to a design that includes a wide-field-of-view camera that shows most of the meeting room at the hub site (see Fig. 3). The satellite views this panorama of the hub room displayed in a window that is full-screen width across his display. Relative to this window, the satellite's camera is positioned horizontally centered and vertically as close as possible.

This view gives the satellite a good sense of the spatial relationships among the people and objects in the meeting room. He can maintain awareness of the locus of attention for each of the hub participants. Because the camera the satellite views is positioned near the screen representing him, he has a good sense of when a hub participant looks directly at him or gestures toward him.

The reverse, however, is not true. The hub participants have a general sense of whether the satellite is looking left, center or right, but nothing more fine-grained than

that. The video mediation introduces too many invisible parameters, such as the field of view of the wide-angle camera and the size of the satellite’s desktop monitor, preventing the hub participants from having a precise sense of the satellite’s focus of attention. With the multiple potential foci of attention in a typical meeting (e.g., people, whiteboard drawings, artifacts), breakdowns occur when the hubs cannot determine the satellite’s focus of attention.

Moreover, the hub participants do not have a visceral sense of mutual eye contact with the satellite participant. When he looks straight at the camera, *all* of the hubs perceive him to be looking straight at each one of them. We call this the *newscaster effect*¹ [5]. When he looks to the left, *none* of the hubs perceive him to be looking directly at any of them, but instead experience him to be looking over their right shoulder. This disruption of eye gaze awareness adds to making it difficult for the hubs to maintain awareness of the satellite’s focus of attention.

A more subtle problem is that the satellite is just not as “present” as his collocated counterparts. We have observed several instances where people are talking in order around the table, e.g., to introduce themselves or report status, and the satellite on his proxy is skipped. We call this the *skip-over effect*. Despite the physical presence of the proxy, we believe there are many factors contributing to this deficit in presence.

We sought to mitigate these problems by physically moving the proxy in response to the satellite’s actions. We began with one degree of motion by putting the display of the proxy on a motorized turntable that is controlled by the satellite participant. We called this a *kinetic proxy*.

To explore how kinetic proxy movement may address the newscaster and skip-over effects, we implemented several forms of motion control, and assessed hub participants’ perceptions of the differences. We performed a lab study of the prototype, comparing a stationary proxy, a kinetic proxy under the satellite’s explicit control by mouse cursor, and a kinetic proxy controlled implicitly by the satellite’s head motion. This is the first study to directly compare alternative ways to project attention and evaluate people’s responses.

The study’s setting was a distributed conference consisting of several collaborative tasks. We used measures of conversation effectiveness such as engagement, naturalness of interaction, cognitive effort and sense of presence.

We found that the kinetic conditions were generally better than the stationary condition, with interesting caveats. For example, screen motion toward one person is more akin to turning one’s back (rather than one’s head) toward someone else. We also found unexpected and previously undescribed benefits and drawbacks to both means of control of the kinetic proxy. Among these is that implicit control generates more incidental proxy motion, which increases the cognitive effort experience by hub participants. These findings suggest designs for future experimentation in kinetic proxies.

2 Background

Researchers of interpersonal interaction recognize that the nuances of body orientation and non-verbal behavior—some consciously controlled, others not—are

¹ What we call the newscaster effect is more commonly referred to as the Mona Lisa effect. We prefer the former in this context, as the affordances of live audio and video viewed on a digital screen are more directly comparable, and are closer to most people’s experiences.

important parts of the messages that people send and receive when communicating in person. Goffman highlighted the difference between the expressions that people *give* and those that they *give off* [6]. The former are verbal signs of the content they want to communicate; the latter are nonverbal and contextual. The former are more conscious; the latter are more subconscious. Hall's study of proxemics formalized that, with cultural differences, the way people orient their bodies toward or away from one another indicates their degree of intent to engage in conversation [7]. Face-to-face is more direct, a 90-degree angle is more casual, and a 180-degree angle is more transitory and disengaged. The design and motion of the kinetic proxy, with its explicit and implicit forms of control, is inspired by these insights.

2.1 Gaze Awareness in Videoconferencing

There is a long series of research prototypes that have tried to improve gaze awareness in videoconferencing. Hydra [8] was a 4-way distributed meeting system that packaged a video camera, small display, microphone, and speaker in self-contained, table-top surrogates dedicated to represent each participant at the meeting. Participants could look toward any of the other participants, and everyone would have an appropriate indication of his or her gaze. A study that compared mediated with same-room conversations found very similar patterns in overall speaking time, speech segment duration, and the distribution of turns taken. Our current study draws upon these same three conversational measures (among others).

More recent videoconferencing studies have also focused on gaze awareness. GAZE-2 [9] used several cameras at each site to capture video of each participant. The system detected which video window was being looked at by tracking each participant's eye movements. Each participant's video view was then shared with the other participants in a rendering of a virtual meeting room. Each video in the virtual room was presented as a flat screen, which was digitally skewed to face the remote participant toward which each participant was looking. These virtual screens rotated in a similar way to our single physical screen.

HP's Halo and Cisco's TelePresence are conference room-scale installations that support individual as well as group meetings. Because all of the participants at each site share the same views of the participants at the other site, correct gaze is not maintained as one moves to different positions within the room. The Virtual Window [10] sought to adjust for motion parallax such as this by tracking participants' head motions and moving a remotely-operated camera to simulate the effect of looking through an opening directly into the remote space.

MultiView [11] preserved gaze and gesture spatial relationships for groups of participants in a two-site (extensible to three) conference. Multiple cameras at each site—one facing each participant—sent multiple video streams to a directional viewscreen at the other site. Positioning was carefully arranged so that every participant at one site had a correct, angle-adjusted view of every participant at the other site, relative to his or her seating locations.

2.2 Sociable Robots

Other research prototypes expressly examine gaze direction in human-robot interaction. Yamazaki *et al.* [12] developed robots with movable heads to support

turn-taking in their communications. These robots engaged humans in one-on-one monologue or simple dialog while orienting their heads toward people or objects of interest. The studies emphasized the coincident timing of robotic gestures with transitional words. Our work also explores how orientation cues can influence interaction, but in a highly collaborative context.

Such robots also act as agents rather than avatars. By representing themselves in an interaction rather than a human other, and by not simultaneously presenting live video of that remote other, they avoid the potential to both complement and contradict an operator's actions. Kinetic proxies take this hybrid approach to providing physical motion as well as onscreen video.

2.3 Kinetic Proxies

A number of embodied telepresence systems have focused on kinetic proxies for hub-and-satellite interactions. PROP [13] was a series of explorations of mobile, robotic personal stand-ins, composed of a video camera and LCD panel (and later, a small pointer) mounted atop a vertical pole and connected to a drivable base. Due to mobility constraints, PROP's primary means of directed gaze was through a pan-tilt-zoom camera head, which served as a partial indicator of the operator's focus of attention, much like our proxies. But as we have found, this can be an ambiguous cue, as the camera may not always follow the operator's attention, or agree with his or her gaze. This overall form and interaction experience has recently appeared in commercial telepresence robots, including Willow Garage's Texai [14], Anybots' QB [15], and InTouch Health's Remote Presence [16].

Sun's Porta-Person [17] prototype also addressed the social presence of remote participants through motion, but specifically within "hybrid meetings," which include a mix of conference rooms, remote and local participants. The device included a video camera and display—replaced by a laptop computer in a later design—stereo speakers and microphones, all mounted atop a turntable and positioned on, or alongside, a conference table. Porta-Person and its turntable represent a direct lineage influence on the physical design of our kinetic proxy.

MeBot [18] was a small, desktop proxy with a three degree-of-freedom head that displayed cropped video of the operator's face, mounted to a mobile base with articulating arms. A study found that the proxy displaying motion was more engaging and likable than without motion. The role of motion as an indicator of attention was not evaluated, and since the participant's head motion was tracked (only), alternative forms of control were not compared.

Though it is not a telepresence proxy, the RoCo prototype [19] is relevant because it uses physical motion to influence engagement. It consisted of an LCD screen mounted on a 5 degree-of-freedom robotic "neck" that could rotate, lean and gesture expressively, mirroring the posture of the person standing in front of it. Studies of RoCo demonstrated that the system could create emotional engagement. RoCo, with its gesture mirroring capability, was an inspiration for our implementation of implicit control.

GestureMan's [20] goal was to support a remote operator in projecting his or her intentions in a workspace shared with a human collaborator. Unlike other proxies, it did not support live video of the operator. Instead, it had the ability to orient its own

robot head, body and a pointing arm, which were controlled by tracking the operator's head movements, screen touches and joystick use.

Animatronic Shader Lamps Avatars [21] were another form of kinetic proxy: a life-scale Styrofoam head mounted on a pan-tilt unit, onto which a video feed of the satellite operator's likeness is projected. The system tracked the operator's yaw and pitch head motions and mirrored them on the avatar. An advantage of this approach was that it presented correct focus of attention cues for all of the individuals interacting with the avatar and over a broad range of viewing angles. It has not been systematically studied from a human-factors perspective.

3 Laboratory Study

To test the effectiveness of the kinetic proxy, we conducted a laboratory study to compare it to a typical stationary video display, and to compare explicit and implicit motion control mechanisms. The study sought to explore the satellite's ability to project gaze cues under these alternative conditions, and hub participants' resulting sense of gaze awareness and presence, by including the directional affordances that people enjoy in face-to-face conversation.

3.1 Study Design

We ran the study as a within-subject design to encourage participants to make comparisons that primarily reflect the absence or presence of motion affordances, and their form of user control. Each group of participants experienced all of the following three conditions:

Stationary: The proxy showed no physical motion at all. This condition most closely resembles a traditional video-chat style conference.

Explicit Control: The proxy screen swiveled in response to the satellite participant explicitly selecting the location she wanted to aim her proxy towards. The position of the proxy was directly linked to the position of the mouse cursor over the panorama view (there was no need to click the mouse button).

Implicit Control: The proxy screen swiveled in response to where the satellite was looking, based on automatic tracking of her head motion.

We set out to test two hypotheses about the perception of motion and control of a kinetic proxy:

Hypothesis 1 (H1): Physical motion of the proxy results in greater conversational engagement, improved sense of directional attention, and preferred interactions by hub participants, compared to no motion at all.

By *physical motion*, we mean the physical movement of the proxy within the meeting room (where the hub participants are located). This is in contrast to apparent motion, which might be represented by repositioning a projected image on a stationary screen. For this study, we focused on physical motion of the screen that displayed the satellite participant's video stream.

Hypothesis 2 (H2): Implicit control of the proxy results in more natural interactions, with lower cognitive effort, and greater sense of the satellite participant's reactions, compared to explicit control.

In both cases, the goal is to more closely reflect the way that people interact during a collocated discussion, or at least, a reported improvement over the non-motion conference experience. We used a combination of behavioral and perceptual measures that are described in more detail below.

Procedure

We ran six groups of subjects through the experiment. We counterbalanced the ordering of conditions across groups. All of the groups were composed of three collocated hub participants who were recruited, plus a confederate acting as the satellite. Participants were led to believe that the confederate was an untrained recruit like them. The same confederate participated in all of the groups, so that the kinetic proxy would be operated in a consistent way throughout the experiment.

Each group worked in each condition for approximately 10-15 minutes. Immediately after each condition, participants individually completed a questionnaire that asked them to rate their experience with that condition. After all three conditions were completed, participants individually rated their preference among the conditions on a questionnaire and then participated in a semi-structured group interview. All sessions lasted approximately one hour, and the entire session was recorded using two overhead cameras and microphones in the hub room that captured the team's activity for later analysis.

Tasks

During each condition, the group performed a decision-making task with no right answer [22] that was intended to evoke discussion and interaction within the group. The following three tasks were always performed in the same order.

Task 1: Decide on a local restaurant to visit as a group after the study (hypothetically) that would work for everyone's dietary constraints and interests.

Task 2: Recommend a number of sites or attractions for a first-time visitor to the region, identified as an acquaintance of the satellite participant.

Task 3: Generate a personalized license plate for a well-known regional celebrity figure, whom the group selected from a short list of alternatives.

Participants were instructed at the beginning of the experiment that the members of the group with the best solution to Task 3, as judged by the experimenter, would receive a \$20 gift card. (In fact all participants received the gift card.)

Participants

The 18 participants (9 male, 9 female) were recruited from the local region and did not know each other prior to the study. They were given a gratuity for their participation. Participants ranged in age from 20 to 55 years old. Their prior experience with videoconferencing varied from this study being their first exposure, to participating in

conferences on a weekly basis. While individual groups had an uneven makeup, every group had both genders (the confederate was female).

3.2 Experiment Setup

Turntable Kinetic Proxy

The satellite participant and hub group were located in adjacent rooms, and communicated through a videoconferencing proxy that we built for the study (see Fig. 2). A 12" Tablet PC supported in the portrait orientation displayed head-and-shoulders video of the satellite participant. The tablet was mounted atop an 8" turntable, which the satellite could remotely position within $\pm 90^\circ$, to directly face any of the hub participants in the room. The frame, constructed of $\frac{1}{4}$ " sheet acrylic to minimize its visual appearance, positioned the display approximately at eye level to the seated hub participants (see Fig. 1). The proxy and hub participants were evenly distributed around a 3' round conference table.

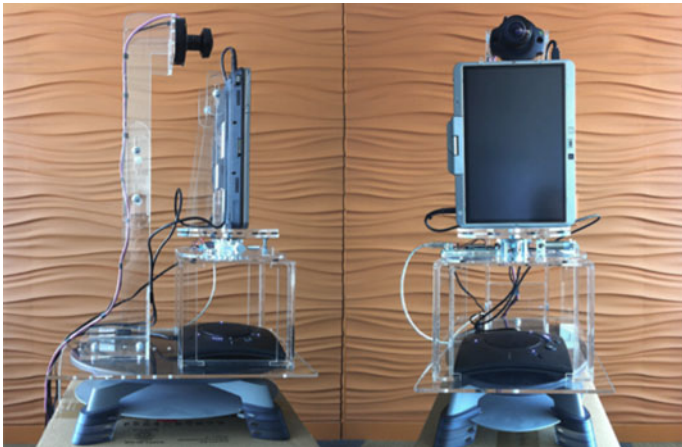


Fig. 2. Side and front views of the kinetic proxy. At its top is a fixed-position wide-angle video camera. Below the camera is a 12" tablet PC, which shows video of the satellite participant. The tablet is mounted atop a remotely-operated turntable which, in turn, is mounted atop a hutch that holds a videoconference speakerphone.

The proxy also included a fixed-position Axis 212 wide-angle camera. The view from this camera was displayed across the entire width of the satellite's 30" monitor (see Fig. 3). This configuration allowed the satellite to see all of the hub participants and their positions around the table, as well as the top edge of the Tablet PC, to confirm that it was oriented as she expected. The satellite's screen also displayed the video directly from the tablet's integrated webcam, but the confederate preferred to focus on the larger, wide-screen image, both to more directly engage the hub individuals and because it provided sufficient feedback of the screen's orientation.

All audio and video communication was over wired and wireless LAN, while the position control stream for controlling the motion of the proxy was carried by USB cable between the two rooms.

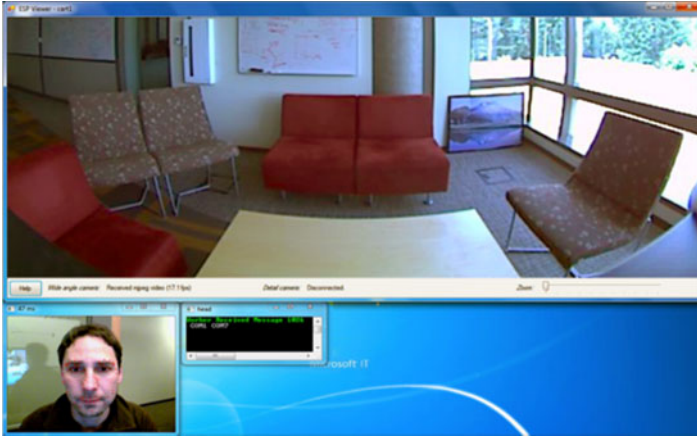


Fig. 3. View of the satellite's interface. At the top is the view coming from the fixed wide-angle camera. The lower left window provides feedback of the image the satellite is projecting on his or her proxy. It also provides feedback that the head-tracker has a good representational model of the satellite's head.

Explicit and Implicit Control

For the explicit control condition, the satellite participant moved her mouse to place the cursor at a particular spot on the client's widescreen view of the hub's workspace. Doing so sent a command which rotated the proxy's screen to face that location in the room. The client program updated the desired 'go-to' position approximately 30 times per second. Since the proxy's turntable was only capable of rotation in the horizontal plane, we only tracked the horizontal component of the cursor's position.

The satellite's wide display meant that she had to turn her head to see the hub participants to her left and right. For the implicit control condition, we tracked this head motion using an in-house, webcam-based software head tracker. This mode also updated the proxy position approximately 30 times per second. Similar to the explicit control condition, we used only the horizontal component of head rotation to orient the proxy's screen.

In both cases, proxy motion was calibrated so that the satellite's mouse or head motions mapped directly to the intended positions around the conference table.

3.3 Measurements

To measure the effects of the experimental conditions, we collected both objective and subjective data during each videoconference session. The objective measures included a tally of responses to deictic prompts and sociometric data that was captured by sensor badges that all of the participants wore around their necks [23].

The subjective measures included the individual questionnaire after each session that focused on the quality of interaction with the satellite member during that condition, the final questionnaire that probed participants' perceptions and preferences about group communication across all three conditions, and the semi-structured group interview, where we discussed these issues in greater detail.

Sociometric Badges

Each participant wore a sociometric badge that consists of several sensors, including an internal microphone, accelerometer, and infrared emitter and detector, which store their data on a removable Micro-SD card. The sensors are all housed in a lightweight black plastic case that is similar in shape, but slightly smaller in size, than a deck of playing cards. One badge was worn around each participant's neck on a lanyard that positioned it about mid-chest.

The microphone recorded the wearer's speech amplitude and time codes [23], and was the only sensor used in this study. Badge data from each day's sessions were downloaded, combined and analyzed using scripts that had been developed earlier. These scripts measured speaking time, speech energy, speech-segment length, and turn-taking per participant and condition.

Five badges were used in total: one worn by each hub participant, one worn by the satellite, and one placed on the proxy. The badge worn by the satellite was not used in the analysis, as the one on the proxy produced a duplicate but cleaner audio signal.

Deictic Prompts

Interspersed through each session, the confederate satellite participant would directly address one of the hub group members with a question of the form "What do *you* think?" or "What type of food do *you* like?" Often, this *first* question was followed with similar *second* or *third* questions to the other two members about their opinions. Responses to these questions were tallied during subsequent video analysis, paying attention to whether it was a first, second, or third question in a series. This distinction is important, because there are three potential respondents to a first question, while there are two for a second question, and only one for a third. Because of this increasing likelihood of responding correctly, we only analyzed responses for the first question according to the following protocol:

Correct Response: The intended person responded immediately.

Correct Confirmation: The intended person checked or confirmed whether he or she was being addressed before responding (e.g., "Do you mean *me*?").

Multiple with Correct: More than one person responded immediately. The group that responded included the intended person.

Incorrect Response: Someone other than the intended person responded immediately.

No Response: None of the participants responded to the question. An example is several seconds of silence, followed by a new thread in the conversation.

Two other codes are possible: 1) someone other than the intended person checked or confirmed whether he or she was being addressed before responding and 2) more than

one person responded immediately but this did not include the intended person. Neither of these categories was present in our data.

The number of deictic prompts varied from session to session, depending on the flow of the conversation or the personality of the participants, but when summed over the study, each condition had 12 first-question prompts and responses.

Questionnaires and Post-Study Semi-Structured Interview

The questionnaire after each condition included nine Likert-style statements and a single brief written response (see Table 1). Each had a 7-point scale with “Strongly Disagree” or “Strongly Agree” as their endpoints. The written response question invited participants to, “Please comment on your experience interacting with the remote person in this session.”

The final questionnaire included three questions that asked which of the three conditions the participant felt the group communicated best with the satellite participant, as well as which condition was most preferred and least preferred.

Following the last of the three conditions, study administrators explained the nature of the study, revealed the confederate’s role in the study, and discussed with the group reflections on the experience. The group interview provided an opportunity to interactively engage participants about their responses, and to follow-up on particular comments. The debrief sessions were video recorded and reviewed after the study.

4 Results

4.1 Conversational Engagement

Our goal was to measure the influence of the kinetic proxy on conversation effectiveness. For this comparison, we included sociometric measures such as speaking time and energy, segment length, and turn-taking. Prior work demonstrates how these measures can characterize interaction [24], as well as establishes desirable values and directionality for the measures in face-to-face interaction [7, 25, 26, 27].

In order to assess participants’ level of engagement in the conversation, we measured their communication behavior using the sociometric badges. A within-subject analysis of this data detected several significant differences between the conditions, which suggest that the different configurations did have an impact on participants’ engagement.

Speaking Time

Greater speaking time suggests a higher level of activity in a conversation [25]. We found a significant difference in the percentage of time that people spoke in each of the conditions ($F_{2,46}=3.62$, $p=0.03$). Participants in the kinetic conditions on average spoke for a larger percentage of time (Explicit: 18.7% and Implicit: 17.8%) than in the Stationary condition (16.6%). Posthoc pairwise comparisons showed a significant difference between the Stationary condition and the Explicit condition ($p=0.02$), and a marginally significant difference between Stationary and Implicit ($p=0.09$), but no difference was detected between Explicit and Implicit ($p=0.29$).

Speech Energy

Speech energy is the variance in speech volume. Higher speech energy often correlates to the perceived excitement of speakers [26] and can be used to indicate their level of activity in a conversation [28]. We found a significant difference in the variance in speech energy in each of the conditions ($F_{2,46}=3.99$, $p=0.03$). Participants spoke with higher speech energy in the Explicit (0.110 units) and Implicit conditions (0.107 units), than the Stationary condition (0.103 units). Posthoc pairwise comparisons showed a significant difference between the Stationary condition and the Explicit condition ($p=0.008$), and between Stationary and Implicit ($p=0.04$), but no difference between Explicit and Implicit ($p=0.44$).

Speech Segment Length

Speech segment length can indicate level of attentiveness and engagement in a conversation [25]. Speech segment lengths are shorter when there are more interjections such as “Oh,” “Uh-huh,” or “Wow” and when there are more frequent turn transitions. More interjections and turn-transitions may show that the listeners are more attentive to or engaged with a main speaker. Hence calculating the average segment length of all types of speech (such as interjections, interruptions, or full turns) allows us to estimate the attentiveness of the conversation. We found a significant difference in the length of speech segments in each of the conditions ($F_{2,46}=10.53$, $p<0.001$). The average speech segment length was longest in the Explicit condition (0.80 sec), followed by the Implicit condition (0.75 sec), while the Stationary condition had the shortest speech segments on average (0.72 sec). Posthoc pairwise comparisons revealed that the Explicit condition had significantly longer speech segments than both the Implicit condition ($p=0.01$) and the Stationary condition ($p<0.001$), but no difference between Implicit and Stationary ($p=0.13$).

Number of Turns per Second

Conversation turn-taking can indicate level of activity in a conversation. The level of interaction among the group members can be estimated by the frequency of turn-taking per second [24]. We found that the number of speech segments per unit of time was significantly different across the conditions ($F_{2,46}=4.8$, $p=0.01$), with the Explicit and Implicit conditions having more turns per second (0.77 turns/sec and 0.75 turns/sec, respectively) than the Stationary condition (0.63 turns/sec). Posthoc pairwise comparisons showed significant differences between Explicit and Stationary ($p=0.009$), and Implicit and Stationary ($p=0.01$), but no difference between Explicit and Implicit ($p=0.72$).

Turn-Taking with the Satellite Participant

We also examined turn-taking in relation to the satellite participant to see if turn-taking to and from the satellite participant was affected by condition. More turn-taking with the satellite participant indicates more active involvement of the satellite participant in the conversation. We found a significant difference ($F_{2,34}=6.4$, $p=0.005$) with hub participants having more conversational turns after or overlapping the satellite participant in the Explicit condition (0.09 turns/sec) and the Implicit condition

(0.08 turns/sec) than in the Stationary condition (0.07 turns/sec). Posthoc pairwise comparisons showed that both Explicit and Implicit had significantly more conversational turns with the satellite participant than the Stationary condition ($p=0.007$ and $p=0.006$, respectively), but no difference between Explicit and Implicit ($p=0.90$).

Relationship between Measures

When compared to the static condition, implicit and explicit conditions were associated with both longer speech segment length and number of turns. For tasks where the goal is to share a fixed amount of information, total speaking time may be somewhat constant among groups, so the number of turns depends on how information is shared in each turn. For open-ended tasks such as ours, total speaking time is highly dependent on how comfortable participants feel, how many ideas come up, etc. Groups could have more turns (to propose ideas) as well as longer speech segments (to elaborate on them). This interpretation is confirmed by the differences in speaking time. Prior studies do not reveal a general trend: Angura [29] shares the negative relationship that we found, whereas Mutlu *et al.* [30] shows a positive relationship.

4.2 Directing Attention

Responding appropriately when being addressed is important for fluid, natural interaction. For the deictic prompts in our study, we examined whether the intended person responded appropriately (see Fig. 4). Both motion conditions (Explicit and Implicit) had the highest number of correct responses, where 100% of the time (twelve instances in each condition), the correct person responded (although in two cases, others in the group also responded, indicating some ambiguity). Additionally, in the Implicit condition, one of the correct responses first sought confirmation. The Stationary condition was the most problematic. In four of the twelve instances, either the incorrect person responded, or no one responded. Examining just the number of correct responses, we found a significant difference across the conditions (Kruskal Wallis Test, $\chi^2=6.049$, $df=2$, $p=0.049$), with Explicit and Implicit having more correct responses than the Stationary condition.

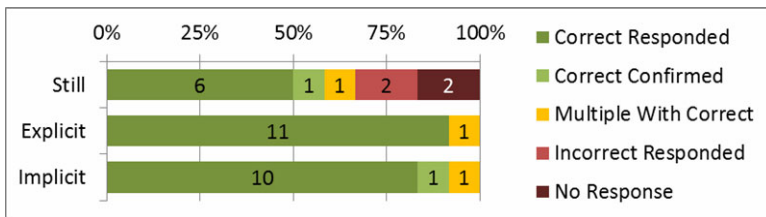


Fig. 4. Results of the confederate’s deictic prompts, as determined by observing the responses in the session videos

4.3 Reactions to the Proxy

Table 1 shows the nine Likert questions participants were asked at the end of each session. The responses revealed that our participants felt positively about their experience

interacting with the satellite, in all of the conditions. They felt that they worked together as a team (average rating 6.6 out of 7) and that they communicated well with the remote person (6.4). They also indicated that it was natural to talk with the remote participant (5.6), and that it was comfortable (6.2) and not fatiguing (6.1) to interact with her.

Table 1. Mean post-task questionnaire results on a normalized* 7-point scale from most negative (1) to most positive (7). * Indicates that the scale has been inverted to match the other questions, where higher numbers indicate positive responses. (⁺p<.05, [±]p<.01)

Question	Stationary	Explicit	Implicit
1. Worked together as a team	6.8	6.4	6.6
2. Communicated with the remote person	6.4	6.2	6.6
3. Interaction with the remote participant was very comfortable	6.2	6.2	6.2
4. Was fatiguing to talk to the remote participant*	6.1	6.0	6.1
5. Felt natural to talk with the remote participant	5.7	5.4	5.6
6. Did not have a good sense of the remote participant’s reactions*	5.8	6.1	5.9
7. Often confused about whom the remote participant was talking to*±	5.1	5.8	5.7
8. Could easily tell when the remote participant was talking directly to me±	4.7	5.8	5.2
9. Sense of the remote participant ‘being there’ ⁺	6.2	5.4	5.9

In terms of confusion about whom the remote participant was talking to, or being able to easily tell whom the remote person was talking to, ratings were lowest (more confusing) in the Stationary condition, and highest (less confusing) in the Explicit and Implicit conditions; however, the differences across conditions were not statistically significant (Friedman Test, $p=0.054$ and $p=0.083$) (see Fig. 5).

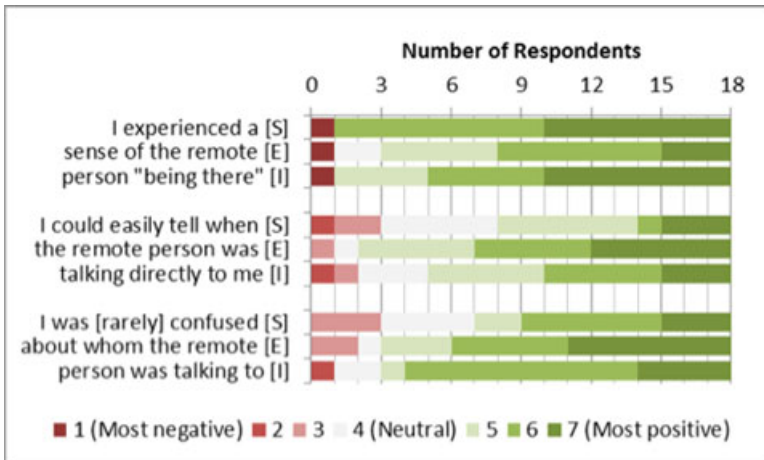


Fig. 5. Results of three post-session 7-point Likert questions: [S] = Stationary condition, [E] = Explicit, and [I] = Implicit. Responses to the third question are presented inverted to match the scale of the other two.

Interestingly, participants rated the sense of presence for the remote participant higher in the Stationary condition (6.2) than the Explicit (5.4) and Implicit (5.9) conditions (Friedman Test, $p=0.003$) (see Fig. 5). Pairwise comparisons showed a significant difference between the Stationary and Explicit conditions ($p<0.032$).

4.4 Overall Preference

At the end of the session, all participants were asked to indicate “In which of the three sessions did you and the group communicate best with the remote participant?” Four indicated the Stationary condition, five indicated the Explicit condition, and seven indicated the Implicit condition (with two stating no preference). Overall, participants felt that the kinetic conditions were more effective by 3:1 over Stationary.

4.5 Satellite’s Perspective

Since the satellite in the lab study was a confederate, she was able to reflect on her experiences across all groups and conditions. She found that in the Explicit condition, she became more consciously aware of who she was directing her attention to than in the Implicit or Stationary conditions. Each movement of the mouse was made in order to either demonstrate listening to a particular person or direct speech toward someone. But over time, intentionally using the mouse in this way became more like a natural extension of her nonverbal communication and movements became more automatic.

In the Implicit condition, she sometimes found herself intentionally “driving” the proxy with her head, rather than moving naturally and trusting the proxy to follow her actions (perhaps due to some technical difficulties with the prototype). She also noticed that, unintentional head movements sometimes distracted the hub participants, making her a bit self-conscious about the way she moved her head.

Overall, the satellite had a slight preference for the Explicit condition. Each movement was meaningful, and participants seemed to pay attention to each movement and interpret its intent correctly. The proxy movement also added communicative value compared to the Stationary condition. She felt that her “voice” was amplified by the motion, and having explicit control over this was the most comfortable for her.

5 Discussion

Our lab study results showed that motion in the kinetic conditions performed better than the Stationary condition in terms of conversational engagement, accurate responses to deictic prompts, and trends in user ranking, confirming H1. A participant illustrated these results with the comment, “The motorized action brought the remote person to life.” Hub participants were able to perceive the satellite’s attention in motion through the swiveling of the display.

However, we also discovered some tradeoffs with motion. The swivel motion of the display could clearly communicate focus of attention (“Rotating LCD made it more clear who remote participant was talking to.”). But, the more general locus of attention suffered, especially because swiveling toward one hub participant often meant that another hub participant was left looking at the edge of the display screen (“...when the remote person was talking to other people, I couldn’t see her and I felt

excluded.”). Given the flat surface of the display screen, swiveling the display was more narrowly directional than the physical affordance of head and body orientation. We expect that these tradeoffs are the main reason why participants rated the motion conditions as lower in a sense of ‘being there.’ Thus there was a tension in motion as it directed attention towards some hub participants but excluded others.

Rotating the display also introduced delay in the conversation, especially when the satellite had to explicitly control the aiming of the display (“...and then kind of like these awkward interruptions, every time she turned.”). Some also found the motion to be distracting (“I felt like whenever she turned we all kind of like stopped talking for a second.”).

Furthermore, aiming the display in the direction of the satellite’s gaze did not address the eye contact issue. Especially in the Implicit condition, where turning the head was used to aim the display, the combination of swiveling the screen and having the head aimed off center from the camera combined to disrupt a true sense of eye contact (“Seemed as if she was looking over my shoulder.”).

Our lab study also illustrated the tradeoffs between explicit and implicit control of motion. Explicit control showed the intent of the satellite more clearly, and was preferred by the satellite confederate, but incurred a delay in operating the interface to aim the display. Some evidence for this delay is found in the longer speech segment measure for the Explicit condition compared to both Implicit and Stationary conditions. This measure may reflect protracted speech while the satellite is simultaneously aiming the direction of the proxy. Or, explicitly aiming the display may leave the satellite’s “gaze” aimed at a conversation partner longer than natural, protracting the partner’s speech turn until the display turns away.

Implicit control tried to lessen the delay in moving the screen and reduce the cognitive burden on the satellite in aiming the display. However, it also added more ambiguity in the intention of the satellite, and the increased amount of motion exposes the hub to more of the negatives of motion (i.e., distraction, noise). Head motion in the physical world can be communicative (e.g., turning toward someone to elicit their response) or incidental (e.g., a side effect of not being able to remain completely still). But the kinetic display motion generated by the turntable was largely interpreted as communicative. Consequently, implicit control caused incidental head movement to be perceived as communicative movement, leading to more of a sense of distraction. In this way, implicit control transferred cognitive effort from the satellite to the hub.

We also expected that hubs would perceive implicit control as more natural than explicit, as more of the satellite’s gestures would be available to them, but we found mixed indications in the questionnaire responses. While our results are equivocal about H2, we have a richer understanding of the tradeoffs between explicit and implicit control.

It is interesting that measures of behavior (sociometric turn-taking, deictic prompt responses) were more demonstrative than perceptual questionnaire responses. Some participants reported not even noticing that the display remained stationary during that condition. The behavioral measures showed that participants reacted to the motion conditions even though their perceptual rankings do not show statistically significant differences. Taken together, these results show that people’s mechanisms of attention awareness may operate at a subconscious level, as has also been seen in other research [31].

6 Conclusion and Future Work

Returning the two practical problems with our proxy that prompted this exploration, do we believe that the kinetic proxy will address the skip-over and newscaster effects? Regarding the skip-over effect, we certainly believe the attention projection provided by motion will give the hub participants enough awareness of the satellite's attention to include her in the conversation. We look forward to deploying kinetic proxies into everyday usage to gain more experience with that. Regarding the newscaster effect, we set out to improve gaze awareness, in the tradition of the research prototypes reviewed earlier that have attempted to do so. Kinetic motion did provide a physical sense of gaze direction. However, it did not achieve eye contact, as the satellite's turning head turned the kinetic proxy display, but also caused her gaze to be directed off angle from the camera.

Our study shows that despite not achieving true eye contact, the kinetic proxy does project the satellite's attention focus so that hub participants could have engaging conversations and correctly respond to deictic requests. Our experiences with the lab study have led us to explore teasing apart attention awareness from gaze awareness.

Eye contact and gaze awareness are mechanisms used for attention projection and awareness when face-to-face. There has been a long series of research prototypes that have indicated how difficult it is to re-create eye contact and correctly convey gaze awareness in videoconferencing. But there are other mechanisms for conveying attention awareness without having to recreate eye contact. While Fels and Weiss [32] have begun to explore this space, we see more opportunities to support attention awareness without relying strictly on eye contact and gaze awareness.

We set out to explore using motion to improve interaction with the satellite participant. We discovered that motion helps, but has some tradeoffs. Swiveling a flat display screen toward one hub participant often excludes other participants, which can diminish the sense of presence of the satellite. Plus, rotating the visual mass of a display incurs lag and some found it to be distracting. Furthermore, swiveling the display did not succeed in improving eye contact.

Based on our study results, we would like to explore designs that leverage the benefits of physical motion, but avoid the exclusion of turning away from participants. Since swiveling the display still did not create true eye contact, perhaps there are ways to use a physical pointer, like a weather vane, to indicate attention projection while keeping the flat display stationary, so all hub participants maintain visual contact with the satellite. Alternatively, it would be interesting to explore convex displays, rather than the flat display screen, which might afford a wider range of directing a satellite's gaze while not 'turning her back' on some participants.

By distinguishing between gaze awareness and attention awareness, what we learned in our lab study generalizes beyond the particular turntable proxy that we examined. The motion of our turntable proxy did provide a stronger sense of attention projection and awareness, even though it did not offer true eye contact.

Complementary to the approaches of re-creating eye contact in videoconferencing systems, we should also explore ways of providing attention projection and awareness which may not depend on gaze awareness. This approach may open up options that are mechanically simpler, more abstract, and perhaps more diverse than previous approaches for creating engagement through videoconferencing solutions.

The current study focused on ameliorating the social asymmetries particular to hub-and-satellite teams. We have had meetings with multiple satellites in attendance by static proxy. In our experience, they have proceeded in much the same way—with comparable improvements in social integration but shortcomings in newscaster and skip-over effects—as meetings with single satellites. As we construct further prototypes, we hope to explore how interaction quality may differ (such as proxy-to-proxy conversations) through the use of multiple kinetic proxies.

References

1. Argyle, M.: *Bodily Communication*. Methuen, New York (1988)
2. Heath, C., Luff, P.: Disembodied conduct: Communication through video in a multi-media office environment. In: *Proc. CHI 1991*, pp. 99–103. ACM Press, New York (1991)
3. Gaver, W.: The affordances of media spaces for collaboration. In: *Proc. CSCW 1992*, pp. 17–24. ACM Press, New York (1992)
4. Venolia, G., Tang, J., Cervantes, R., Bly, S., Robertson, G., Lee, B., Inkpen, K.: Embodied social proxy: Mediating interpersonal connection in hub-and-satellite teams. In: *Proc. CHI 2010*, pp. 1049–1058. ACM Press, New York (2010)
5. Vishwanath, D., Girshick, A., Banks, M.: Why pictures look right when viewed from the wrong place. *Nature Neuroscience* 8, 1401–1410 (2005)
6. Goffman, E.: *The Presentation of Self in Everyday Life*. Doubleday Anchor, Garden City (1959)
7. Hall, E.: A system for the notation of proxemic behavior. *American Anthropologist* 65, 1003–1026 (1963)
8. Sellen, A.: Speech patterns in video-mediated conversations. In: *Proc. CHI 1992*, pp. 49–59. ACM Press, New York (1992)
9. Vertegaal, R., Weevers, I., Sohn, C., Cheung, C.: Gaze-2: Conveying eye contact in group video conferencing using eye-controlled camera detection. In: *Proc. CHI 2003*, pp. 521–528. ACM Press, New York (2003)
10. Gaver, W., Smets, G., Overbeeke, K.: A virtual window on media space. In: *Proc. CHI 1995*, pp. 257–264. ACM Press, New York (1995)
11. Nguyen, D., Canny, J.: MultiView: Spatially faithful group video conferencing. In: *Proc. CHI 2005*, pp. 799–808. ACM Press, New York (2005)
12. Yamazaki, K., Yamazaki, A., Okada, M., Kuno, Y., Kobayashi, Y., Hoshi, Y., Pitsch, K., Luff, P., von Lehn, D., Heath, C.: Revealing Gauguin: Engaging visitors in robot guide’s explanation in an art museum. In: *Proc. CHI 2009*, pp. 1437–1446. ACM Press, New York (2009)
13. Paulos, E., Canny, J.: PRoP: Personal roving presence. In: *Proc. CHI 1998*, pp. 296–303. ACM Press, New York (1998)
14. Lee, M., Takayama, L.: Now, I have a body: Uses and social norms for mobile remote presence in the workspace. In: *Proc CHI 2011*. ACM Press, New York (2011)
15. Anybots, <http://www.anybots.com/>
16. InTouch Health, <http://www.intouchhealth.com/>
17. Yankelovich, N., Simpson, N., Kaplan, J., Provino, J.: Porta-Person: Telepresence for the connected conference room. In: *Ext. Abstracts CHI 2007*, pp. 2789–2794. ACM Press, New York (2007)
18. Adalgeirsson, S., Breazeal, C.: MeBot: A robotic platform for socially embodied telepresence. In: *Proc. HRI 2010*, pp. 15–22. ACM Press, New York (2010)

19. Breazeal, C., Wang, A., Picard, R.: Experiments with a robotic computer: Body, affect and cognition interactions. In: Proc. HRI 2007, pp. 153–160. ACM Press, New York (2007)
20. Kuzuoka, H., Kosaka, J., Yamazaki, K., Suga, Y., Suga, Y., Yamazaki, A., Luff, P., Heath, C.: Mediating dual ecologies. In: Proc. CSCW 2004, pp. 477–486. ACM Press, New York (2004)
21. Lincoln, P., Welch, G., Nashel, A., Ilie, A., State, A., Fuchs, H.: Animatronic shader lamps avatars. In: Proc. ISMAR 2009, pp. 423–432. IEEE, Los Alamitos (2009)
22. McGrath, J.E.: Groups: Interaction and Performance. Prentice-Hall, Inc., Englewood Cliffs (1984)
23. Olguin Olguin, D., Waber, B., Kim, T., Mohan, A., Ara, K., Pentland, A.: Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics* 39(1), 43–55 (2009)
24. Kim, T., Chang, A., Holland, L., Pentland, A.: Meeting mediator: Enhancing group collaboration using sociometric feedback. In: Proc. CSCW 2008, pp. 457–466. ACM Press, New York (2008)
25. Curhan, J., Pentland, A.: Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology* 92, 802–811 (2007)
26. Hung, H., Gatica-Perez, D.: Estimating cohesion in small groups using audio-visual non-verbal behavior. *Transactions on Multimedia* 6(12), 563–575 (2010)
27. O’Conaill, B., Whittaker, S., Wilbur, S.: Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human Computer Interaction* 8(4), 389–428 (1993)
28. Vertegaal, R., Ding, Y.: Explaining effects of eye gaze on mediated group conversations: Amount or synchronization? In: Proc. CSCW 2002, pp. 41–48. ACM Press, New York (2002)
29. Angura, X.: Robust speaker diarization for meetings. PhD dissertation, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona (2006), <http://xavieranguera.com/phdthesis/node47.html>
30. Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., Hagita, N.: Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In: Proc. HRI 2009, pp. 61–68. ACM Press, New York (2009)
31. Sato, W., Okada, T., Toichi, M.: Attentional shift by gaze is triggered without awareness. *Experimental Brain Research* 183(1), 87–94 (2007)
32. Fels, D., Weiss, T.: Toward determining an attention getting device for improving interaction during video-mediated communication. *Computers in Human Behaviour* 16(2), 99–122 (2000)

Making Sense of Communication Associated with Artifacts during Early Design Activity

Moushumi Sharmin and Brian P. Bailey

University of Illinois at Urbana-Champaign, 201 N Goodwin Avenue, Urbana, IL 61801, USA
{sharmin2, bpbailey}@illinois.edu

Abstract. Communication associated with artifacts serves a critical role in the creation, refinement, and selection of conceptual ideas. Despite the close relationship between ideas and surrounding communication, effective integration of these two types of design materials are not well-supported by existing design tools - resulting in ad-hoc and ineffective strategies for managing communication during the design process. In this paper, we report the results of a contextual inquiry (N=15) aimed at understanding communication practices, its role in the design process, and strategies utilized by designers to manage and utilize communication outcomes in relation to artifacts. Our findings show that more than 50% of early design activity consists of three categories of communication (information seeking, brainstorming, and feedback) and communication practice varies as a function of expertise, organizational and social factors. Additionally, novice and freelance designers exhibit greater reliance on online forums to find suitable communication partners to generate and refine ideas whereas experts communicate with other experts or team members for information collection and sharing.

Keywords: Design, Artifacts, Communication, Ideation, User Study.

1 Introduction

Communication is a central part of the ideation process, especially for creative, non-routine design activity [1, 2]. Many types of communication (e.g., client correspondence, brainstorm discussions, and feedback on initial ideas) influence the ideation process. Research shows that discussion associated with artifacts promotes design thinking [1], enables generation and refinement of alternatives [3], and contextualizes the design process [4]. In this paper, we use the term *ideation process* to refer to activities associated with the generation and evaluation of design artifacts during the early design phases. In addition, we use the term *design communication* to refer to any type of exchange, verbal (face-to-face, over phone) or written (e-mail, IM, etc.), synchronous (phone) or asynchronous (e-mail) related to ideation activities.

Research on design has underscored the importance of communication in the design process [27], examining it in specific situations such as studying communication roles [8] and negotiation strategies [2, 8] within co-located teams [9, 27] or specific types of communication such as rationale related to design choices [10, 11]. Research on collaborative and participatory design focused on the importance of

communication among stakeholders [27]. Research aiming to support early design activities has concentrated on how designers collect, construct, and manage artifacts [6, 7], but has paid less attention to the inter-connection between artifacts and communication. Little emphasis has been given to the integration of artifacts and communication in design tools, forcing designers to invest significant effort to capture and access the connections between them. Research on design reuse has also identified this lack of connection as one of the main barriers for supporting reuse [19].

In this paper, we report results from a contextual inquiry investigating the role of communication surrounding artifacts during the ideation process. We examine details of designers' communication practice, focusing on communication partners, when in the design process communication occurs, the communication channels utilized, and how the outcomes of communication are captured and utilized in the design process. We also probe social and organizational factors that influence designers' communication behavior. Based on our lessons, we propose guidelines to better integrate communication within existing and future design tools.

Our study consisted of semi-structured interviews with professional designers in the creative design domains. We selected designers having different design backgrounds (graphic and industrial) and working in different settings (design firm and freelancers) to identify similarities and differences in their communication practices. We also studied designers having varying levels of expertise (between 1 to 33 years). Our findings show that communication accounts for more than 50% of ideation activity. Designers utilize communication to seek information, generate, refine, and select ideas, and to resolve creative blocks. While work setting and expertise influence designers' communication practice, all designers struggled to capture and connect communication outcomes with relevant artifacts. The main contribution of this work is providing new understanding of communication practices related to the use of artifacts across domains, work settings, and levels of expertise and a set of actionable implications for designing better design support tools.

2 Related Work

2.1 Studies of Design Communication

The role of communication is well researched in the realm of collaborative and distributed work [12, 13, 14]. Kraut et al. studied informal communication as a means for improving collaboration and coordination in an organizational setting [13], while Chiu examined the effect of team organization on communication [12]. Sonnenwald examined communication as a collaboration tool for multi-disciplinary teams [2]. This thread of research focused on communication as a means for improving collaboration efficiency. During the ideation process, the role of communication transcends collaboration efficiency, as it allows designers to learn about the design space and guides the generation, selection, and refinement of artifacts. Communication during ideation often includes artifacts such as sketches, drawings, prototypes, and other visuals [1]. Studies focusing on communication therefore offer only a partial picture of the ideation process, ignoring the influence of communication outcomes in the generation and refinement of artifacts. One consequence is that existing tools do not provide adequate

support for integration of communication and related artifacts; making it difficult to utilize them during ideation. We aim to fill this gap.

Design researchers have attended to formal and organized communication (e.g., brainstorming sessions) [15,16]. Eckert and Stacey conducted several studies investigating communication in collaborative design tasks and offered six broad dimensions of communication, including form of communication, task, subject and tool expertise, and objectives for communication [15]. Stempfle and Badke-Schaub focused on design thinking as a team process and viewed communication as a representation of individual and team thinking [16]. Our research extends prior work by probing the various underlying reasons that motivate and influence communication associated with artifacts in practice.

2.2 Studies of Early Stages of Creative Design Practice

Research on early phases of creative design has focused on different types of artifacts and how these artifacts (e.g., sketches and prototypes) impact ideation process [17, 18]. Bonnardel studied the influence of information availability on the idea generation process[4]. Researchers also studied the impact of capturing, accessing, and reusing artifacts in the ideation process[19]. A separate thread of research centered on the creation and utilization of sketches and early prototypes to externalize ideas[7, 20]. Researchers also focused on facilitating ideation process by capturing the history of the progression of artifacts [21]. Additionally, researchers have focused on the influence of workspace activities within design teams during the artifact generation process [27].

Prior research on early stages of creative design has emphasized the construction of artifacts, especially through sketching, without regard for the rich communication that influenced and was prompted by those artifacts. Researchers consider communication to be an inseparable part of the ideation process [27] as it not only guides the design process, but also allows designers to represent and support their design decisions [1]. Studying artifacts provides only a partial understanding of the early design process, as it fails to portray thought processes, discussions that prompted changes in artifacts, and decisions that led to the evolution of the ideas. We aim to offer better understanding of the relationship between artifacts and related communication by studying designers' communication practices.

2.3 Systems Supporting Design Communication

Most communication support systems focus on capturing decisions and deliberations related to design problems and are aimed to support redesign and maintenance issues [10, 22, 23]. Another related class of system supports communication among stakeholders largely by capturing decisions surrounding artifact design [24, 25]. Mood boards and similar systems often capture design scenarios, explored alternatives, and early design activity to convey ideas and design choices to clients [5,26]. These systems try to provide awareness of ongoing design activity by capturing and representing artifacts or specific types of communication (e.g. rationale), but ignore the relationship between artifacts and other types of communication (e.g., discussions, feedback, and client preferences) that shapes the ideation process. The resulting systems thus fail to represent the design process, making it difficult for designers to

utilize and understand the process in which the artifacts were created[19]. We believe research aimed at understanding designers' communication practice in relation to artifacts can provide valued insight and guide the design of tools that would provide more effective management of the design process.

3 Methodology

The goal of our study was to understand communication practices during ideation activities holistically, investigate designers' motivations and strategies for initiating, managing, and utilizing communication outcomes, and examine organizational and social factors that influence their communication behavior. Our study consisted of semi-structured interviews with professional designers with varying levels of expertise and coming from different creative design domains. We used self-report data to classify designers as expert and novice. For example, one designer has a total of three years of professional experience and rated his expertise level as "novice." Another designer has more than 10 years of professional design experience and rated herself as an "expert." We also considered two types of work setting, designers who work in an organizational setting (design firms, N=8) and designers who work independently (freelance, N=7). See table 1 for background of the study participants.

The interview had 19 questions guided by our prior research on reuse of existing design knowledge during the early stages of creative design domains[19]. Table 2 lists sample questions asked during the interview. We began the interview by asking the designer to briefly describe a recent or ongoing project and encouraged her to share stories about communication practices during their ideation process to ground the discussion. For example, one designer was working on the design of a social networking site for a corporate setting. During the interview, the designer focused on how communication with clients helped to accumulate various types of information needed for designing the site and how discussion with other designers helped to refine his early ideas.

Table 1. Background of study participants

Domain	Number of Participant	Work Setting		Level of Expertise	
		Design Organization	Freelance	Expert	Novice
Graphic	7 (3 female)	5	2	4	3
Web	4	1	3	2	2
Industrial	4 (2 female)	2	2	2	2

Interviews lasted no more than two hours and interviewees received \$20 for their participation. Ten of these interviews were conducted in the designers' workspace and the remaining via phone due to distance. While phone interviews limited our ability to directly examine the workspace, we requested the interviewees to share additional materials such as photos of the workspace, screen-captures of the artifact and communication management systems, and records of captured communication with us to gain better understanding of their communication practice.

Table 2. Sample of the questions asked during an interview

Types of Communication
What are the different types of communication that occur throughout the design process?
Frequency and Motivation
How often do you communicate with other designers during the early design phase? What motivates you to do so?
Capture and Access
Why and how do you capture and access communications for an ongoing or a prior project, if at all? What motivates you to do so?
Medium of Communication
Which is your preferred method for communication during early design (face-to-face, e-mail, etc.)? Why do you prefer this method over the other methods?

All of the interviews were audio-recorded and later transcribed. Each of the transcripts were coded separately and later analyzed for recurring themes across interviews. We also collected relevant artifacts such as stored e-mails, IMs, and files containing traces of communication utilized during the ideation process and analyzed these to verify opinions expressed during the interviews and designers’ actual practices. Though the study was qualitative in nature, numbers were reported to highlight the relative significance of an observed behavior.

4 Study Results

We report the results from our study of communication surrounding artifacts during ideation activities. We discuss why, with whom and when communication occurs, discuss communication categories, channels utilized, and strategies adopted by the designers for managing communication during ideation.

4.1 Role of Communication in Ideation Process

Communication Accounts for more than 50% of Ideation Activity: Designers deem communication as a core design activity and reported that on an average more than 50% of early design activity involves communication related to artifacts. Designers communicate to learn about existing artifacts, generate ideas, compare alternatives, and refine and select ideas for further consideration. They engage in communication about the artifacts and utilize the outcomes of communication to guide the design process. Designers leverage artifacts and associated communication to capture the underlying design process and believe that artifacts or communication alone fails to adequately represent their process. As a result, designers want to capture artifacts and related communication together as it facilitates reuse, in line with findings reported in [19]. Existing technology do not support capturing and management of artifacts and communication in one space, resulting in disassociation between these two design components. This forces designers to come up with ad-hoc strategies to link artifacts and communication for effective utilization.

Though designers mentioned engaging in communication throughout the design process; they believed the impact of communication is most significant during the early phases when they are striving to come up with conceptual ideas. We observed designers' workspaces showcasing different types of early design materials, where communication is captured and highlighted in terms of feedback on post-its, annotation on sketches, list of requirements, etc. Figure 1 presents one such workspace where artifacts and related communication are displayed in the idea space to guide the design process. Idea spaces showcasing artifacts and associated communication is typical in all the designers' physical workspaces, indicating a need for better integration of these two design components in the design support tools.

Communication Helps Resolve “Designer’s Block”: Fourteen out of fifteen designers interviewed mentioned engaging in communication as a method of getting past “designer’s block.” Designers consider communication not only necessary but also as a fail-safe method for removing the block. To quote two designers:

“There’s a point when you reach your road block in the design process and you need fresh perspective, you need someone to ask “hey, what do you think?” That happens a lot.” [P4]

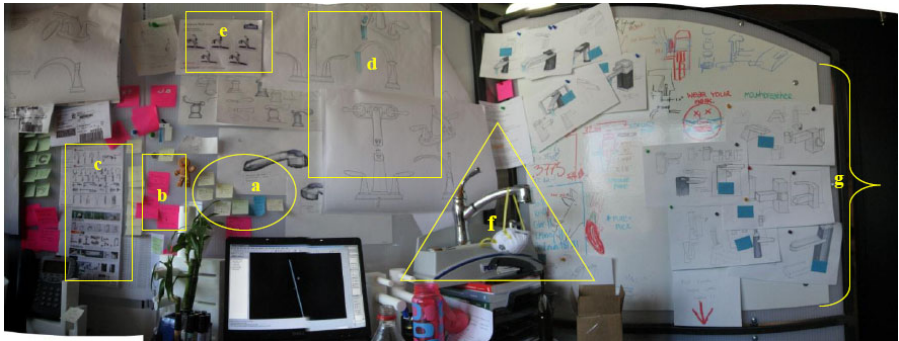


Fig. 1. Different types of early design materials: (a) ideas on post-its; (b) feedback on ideas; (c) materials collected from the Web; (d) prototype sketches; (e) product images; (f) physical products; and (g) notes from brainstorming sessions; displayed in a designers' workspace. Artifacts and communication are captured in close physical proximity to assist in ideation.

“Everyone gets stuck no matter how creative or how good you are. And that’s when other designers will come out and they might be mean about it, but I like to think that everyone won’t be blunt about it. I’ll post (in Web forums) to get help.” [P10]

4.2 Communication in Different Phases of Early Design Activity

Information Seeking – Personal Communication is Preferable to Web Search:

The first step during ideation is to collect information for problem formulation. Designers communicate with other designers, clients, and end users to better situate the design problem, to collect information on the product and similar projects, and to collect requirements from clients. Designers also collect information from office re-

pository, the Web, books and magazines. While Web and other sources allow quick access to information, opportunity to learn from other designers' experience makes personal communication the most preferred method for collecting information. Designers search technical information on the Web, but communicate with other designers for collecting subjective information (e.g., feedback, rationale). See figure 2 for most utilized sources of information during ideation activities. Designers also communicate with other designers whom they believe to be invaluable sources of design knowledge, which can't be attained from any other sources. This preference is exemplified by the following two quotes:

“Usually I'll gravitate towards who I know first. I'll start from my old colleagues, I'll ask my whole group, and then do a search on the Internet. If I am looking for some technical information, I'll first search in the Internet, but if I get really stuck, I'll ask people more about it.” [P8]

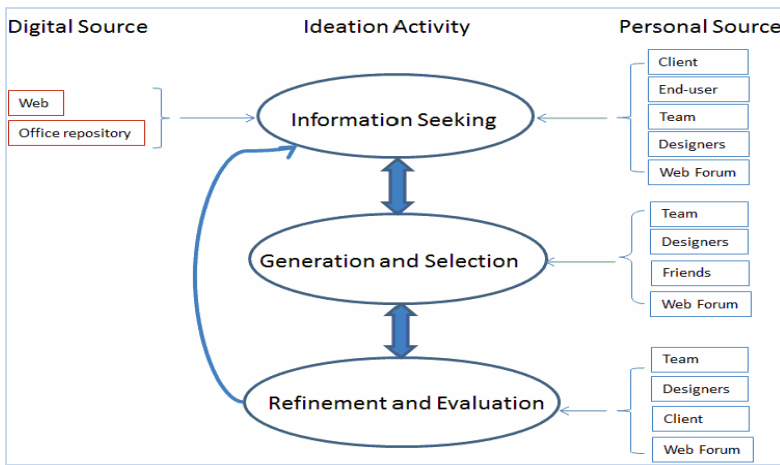


Fig. 2. Information sources utilized during ideation activities. Sources are listed in order of preference (client is preferred than end-users for information seeking).

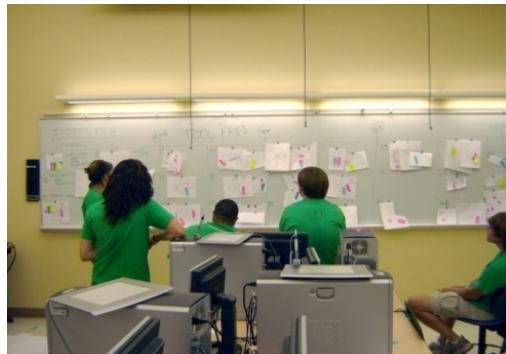


Fig.3. Group communication session targeted to share information and generate ideas. Early ideas are posted on a white board for discussion and selection.

“There is so much knowledge on gaming culture and you can't find it in Google or you can read so many books, but still not find the answer. You have 100~200 people who have these small pockets of knowledge and together it's pretty formidable.” [P7]

Generation and Selection of Ideas – Integration of Artifacts and Related Communication is Imperative: During ideation, designers mentioned engaging in frequent communication with others designers, especially when they get stuck and need fresh perspectives. After initial research, designers engage in brainstorming sessions to share information, to narrow down the problem space, and to generate, refine, and select ideas for refinement. Research on early design also highlighted the importance of team meetings and brainstorming on the innovation process [3].

Figure 3 depicts one such session where members of a group are engaged in ideation. While these sessions indicate close relationship between artifacts and communication, existing tools support capturing either artifacts or communication. As a result, relationships between artifacts and communications are lost. However, we observed designers making an effort to preserve these relationships by capturing the outcomes of communications using notes and e-mails to themselves and storing these along with related artifacts. When asked, designers mentioned that artifacts and communications separately fail to capture the process and rationale for generating and selecting ideas, and the value lies in capturing these two components together. Eleven out of the fifteen designers mentioned creating notes and/or taking pictures of the white boards filled with discussion points to preserve outcomes of the communication with the other design materials. This indicates a need for tools that would better connect communication outcomes with the other design materials.

Refinement and Evaluation through Feedback: Another type of communication prevalent in early design is feedback, which allows designers to refine, improve and validate their ideas. A significant part of early design communication is targeted to receive and provide feedback. Designers also utilize feedback to learn about potential weakness and strength of their ideas and to examine and enhance the quality of ideas. The most cited reason for feedback request is to receive a fresh and alternative perspective on an idea. One designer summarized this behavior nicely:

“If I'm on my own doing a project I get too close to it to see other issues or problems. If someone has a fresh look at it they would say does that really work or how does that work. I would never think of that. Someone hasn't been staring at the same thing. Play devil's advocate. (And ask) Would that really work?” [P3]

All of the designers participating in our study mentioned requesting feedback on their ideas, two of the fifteen mentioned refining the ideas before requesting any feedback, and four mentioned requesting feedback throughout the ideation process.

4.3 Medium of Communication – A Shift towards Digital Channels

Communication occurs throughout the ideation process; it begins even before designers start generating ideas. Designers utilize different communication channels (face-to-face conversations, over phone, via e-mail and IM) based on the underlying design activity and the phase of the design. Figure 4 presents a distribution of utilized and preferred communication channel - indicating a shift towards the digital channels.

While face-to-face communication is considered imperative, digital channels, especially e-mail is the most utilized channel to initiate, manage, and create a record of communication outcomes. Interestingly, 60% of the designers mentioned following up with e-mails after face-to-face or phone conversations. The capability of including images of artifacts along with the discussion and ease of sharing and access make e-mail their preferred choice for communication. One designer stated:

“I hang on to all the e-mails. I actually try to take notes on what people say. I take notes in text files and keep them in the same folder like the Photoshop files.” [P9]

4.4 Communication Capture and Management

Communication Capture: Designers make genuine efforts to capture outcomes of Communication for a number of reasons. Communication in terms of design decisions, feedback, and thoughts on ideas allows designers to present their rationale to

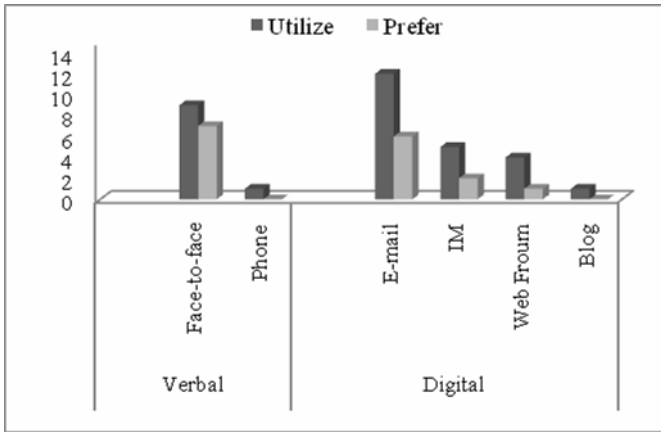


Fig. 4. Distribution of communication channel use. X-axis represents channels and Y-axis represents number of designers. All but one participant reported utilizing multiple channels.



Fig. 5. Anidea space of one designer. Post-its contain annotated sketches displaying comments, feedback, and rationale for each design choice.

the clients and others. Designers consider capturing and presenting feedback on an idea and rationale for selecting a direction as a way of validating their design. To quote a designer:

“When somebody comes and you show them your final product, and they just think that you put those things together somehow magically but actually there were lot of iteration, there's lot of thinking that went behind it, and it's kind of a way to show – “here's actually what I was thinking.”[P4]

Designers also try to capture communication outcomes to select the best idea from a number of possible candidates. To some designers selecting the best direction is often as challenging as coming up with numerous ideas [19]. Designers tend to capture early ideas and associated feedback on post-its and use these to analyze alternate directions. Figure 5 shows idea space for a designer where she accumulated her ideas along with feedback received and rationale behind her design choices. She utilized this space not only for selecting a direction but also for showing the client and others her process. Spaces such as this are common in the design workspaces, illustrating the need for better integration of artifacts and relevant communication.

Communication Management: Existing technology poses challenges in the way designers try to manage communication surrounding artifacts. Designers want to record these communications not only for using in the ongoing project, but also as a reference for future projects. Especially, communication with clients in terms of requirements, feedback, and their preferences are stored for potential future projects. Figure 6 presents one such text file that one designer (P9) created to record all the feedback received for an ongoing project. This file is used to keep track of the suggestions from other designers and his reaction to these suggestions. However, the designer expressed frustration as these communications are disconnected from each other and from the artifacts and he has to invest a great deal of effort to create a single space to combine the outcomes of communication and resulting artifacts.

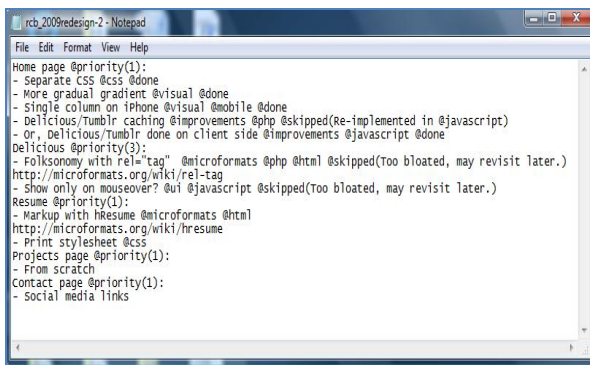


Fig. 6. Combining summaries of feedback received from various communication channels in a single text file

Artifacts are typically stored in directory-based file management systems while communication is stored as e-mail archives, chat logs, voice mails, etc. The separation of these two types of design materials forces designers to adopt ad-hoc strategies to connect them together. For example, we observed designers investing lots of effort to record all the communication related to an artifact by creating files that combine summaries of communications received from all different channels. Another popular strategy was to mimic the directory structure in e-mail clients to reduce the burden of management (see figure 7). Similar finding has been reported in [25] for organizing materials in the area of software design.

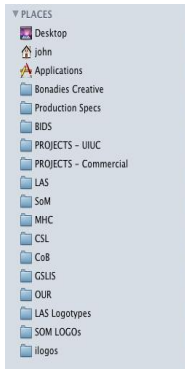
4.5 Influence of Work Setting: Access, Culture, Ownership and Criticism

Work setting influences designers' strategies and motivations for communication. Culture of organization, ownership of ideas, fear of criticism, and access to design resources greatly impact designers' experience surrounding communication.

Access to Communication Partners: Designers working in design firms tend to communicate more during the ideation process as they have access to team members and other designers working in the organization. Freelance designers only have access to clients as most of them work alone. Lack of access to other designers and intellectual property issues force them to limit their communication within a selected and trusted group of designers and with clients. Freelance designers often mentioned relying on close friends and family members to bounce ideas off of due to a lack of access to other experienced designers. One freelance designer mentioned consulting his spouse (a non-designer) when faced with challenge during the ideation process. Freelance designers try to utilize online design forums to gain access to other designers who may eventually become trusted communication partners. Every freelance designer mentioned this lack of connection with other designers as one of the key challenges faced during early design phases.

Figure 8 reflects designers' preference for feedback request. For designers working in organizations, team members are the most preferred source (63%), while freelancers mentioned clients (57%) as their primary source for feedback. This difference is due to the accessibility to other designers, clearly indicated by the fact that designers working in organizations never mentioned clients as a potential source for feedback on early artifacts. However, a fairly large number of designers (43% and 50% respectively) prefer to communicate with friends or peers (who are not involved in the project) about artifacts. A significant percentage of freelance designers (29%, mostly novice) preferred online forums, while none of the designers working in firms considered online forums as useful source for feedback.

Organizational Culture: Organizational culture has been regarded as the primary factor that hinders or promotes communication. Team boundary often discourages communication within an organizational setting as designers feel "less comfortable" communicating and providing feedback to other designers who belong to other design teams. One designer shared his experience with different organizational culture and how it impacted his communication behavior:



(a) Artifact Space



(b) Communication Space

Fig. 7. Managing artifact space and communication space by creating similar structure in the hard drive and e-mail client

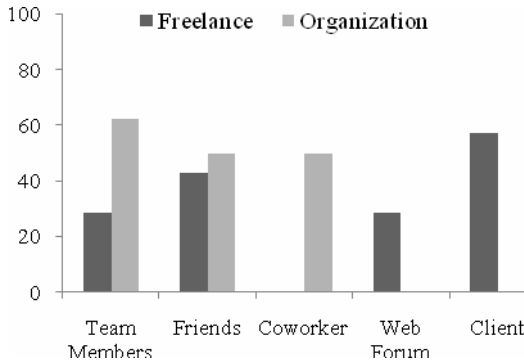


Fig. 8. Designers’ preferred group for requesting feedback. Y-axis represents designers’ preference (%). Empty bars indicate there was no preference for the group.

“Many places are hyper competitive and success mattered on how often you won, and mostly jobs will be competitive among a number of people and it was important that you won, and in some places the culture is such that in a way it rewards that behavior. So people stole ideas and it became a territorial thing, in that case you wouldn't share because you couldn't count on the goodwill and good faith...But here the culture is not like that at all as they don't reward behaviors like that. We are all quite comfortable to walk in our buddy’s office and share ideas and ask feedback.” [P14]

Ownership of Ideas and Fear of Criticism: A large number of designers refrain from communicating early ideas within the organization as they fear that negative or blunt criticism may diminish their enthusiasm for the idea. Competition among team members is also mentioned as one of the reasons why designers feel discouraged to communicate. Even designers working as part of a team tend to delay communication

until they feel comfortable with the quality of generated artifacts. This behavior is exemplified by the following two quotes:

“Typically you don't communicate as much when you haven't got your idea yet. You don't want to sense that you are stealing from someone else's ideas. You don't want to talk to somebody so that they come up with the idea that you yourself don't feel like it's truly yours. These are the times that I am hesitant to get feedback.” [P14]

“I feel that people might criticize things that I am already aware of that need to be changed, so I want to get to a state where I feel like I have done most of, I have contributed most of what I want to do, and then see what other people think about it as opposed to starting something and ask other people before I am done.” [P9]

Majority of the freelance designers use online forums for communicating ideas and expressed that lack of control over the audience and the likelihood of receiving negative feedback are their biggest concerns. One designer went as far as boycotting a forum as he considered the feedback posted as negative and discouraging. This behavior is quite common among freelance designers and can be better exemplified by a comment made by P6:

“A place like site2 (pseudonym) forum will discourage me to share information, I think there are some fantastic designers there but between site2 and site1, it is a very separate group of people when it comes to personalities. (In site2) They do not hesitate to say that you are an idiot or stupid while in site1 it's little bit more laid back, little bit softer, and it's not that harsh.” [P6]

4.6 Influence of Expertise: Communication Categories, Partners and Motive

Level of expertise influences how and why designers communicate, with whom they communicate, and how do they capture and utilize communication in their design.

Communication Categories: While both expert and novice designers mentioned information collection as one of the primary reasons for initiating communication, the types of information that designers request vary depending on designers' level of expertise. Expert designers rarely request technical information from other designers while novices mentioned other designers as the best resource for finding technical information. Expert designers feel socio-psychological pressure of “not knowing enough” and are reluctant to ask for help on technical issues but novice designers consider themselves as learners and feel that they are expected to ask. Expert designers, contrarily, request more for client and product related information to learn more about the design space and from the experience of other designers. Expert designers mentioned that more than 80% of their communication consists of feedback and discussion about design decisions. On the contrary, a significant part of early communication for novice designers is aimed for generating, comparing, and refining ideas along with reputation building. Figure 9 shows designers' motivations for communication. While both expert and novice designers mentioned communication as a means of collecting information, receiving feedback and reciprocity, novice designers communicate more to receive creative input on ideas than expert designers.

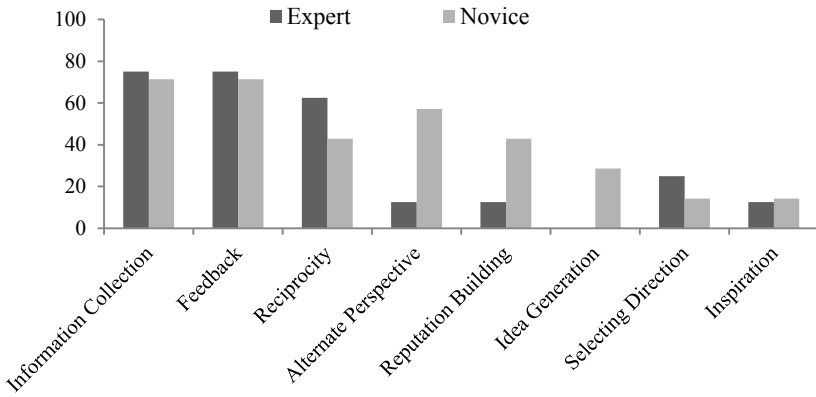


Fig. 9. (Expert vs. Novice) Designers’ motivation for engaging in communication. X-axis represents reasons and Y-axis represents the percent of designers who prefer them.

Communication Partners: Finding appropriate communication partners often poses a problem for novice designers. Trust plays a significant role in selecting designers for communication about early ideas, often due to fear of negative criticism and threat of exposing unfinished ideas. While expert designers utilize prior experience to find trusted partners, novice designers often struggle in this process. Novice designers go as far as to study the history of communication of designers (collecting information from colleagues or reviewing past communication) to select the best candidates for communication. Novice designers try to build relationships by posting on design forums and by contacting designers personally. However, these approaches help little as designers rarely afford to invest time to analyze communication history of individual designers to find a proper fit. Automated recommender systems can greatly help in the process of finding suitable communication partners.

Utilization of Communication Outcomes: Expert and novice designers’ tend to communicate their ideas at different times within the ideation process. Novice designers tend to communicate throughout the ideation process and use the feedback to refine ideas. Expert designers rarely share their initial ideas, waiting till it reaches a certain point where they are either satisfied with the idea or are suffering from “designer’s block.” Novice designers also show more openness towards harsh feedback as they view it as learning process. To quote a designer:

“Experience plays a role (in how you utilize it). If you are young, you ask feedback differently and receive it differently, and you would accept even challenging or disappointing feedback. And you also don’t have the confidence to say that I hear you, but I don’t buy it. The younger you are the more likely you are to do everything anybody tells you. And having more experience, you learn to understand what is important and what might be ok but not certainly important.” [P14]

5 Discussion and Design Implications

In this section we discuss implications derived from lessons from this study. The implications presented are not exhaustive, but they offer guidance for better connecting communication and relevant artifacts and removing barriers for engaging in communication for ideation.

5.1 Integrating Communication Traces with Artifacts

Artifacts and surrounding communications are the two main components of ideation process. Communication traces help designers recall their design choices and convey their process and rationale to others. However, the disassociation of artifacts from communication makes it challenging to effectively utilize communication outcomes during the ideation process. Lack of suitable mechanisms for connecting communication and artifacts for long term management makes it extremely difficult to access, retrieve and utilize communication outcomes from ongoing or past projects. Every designer studied expressed frustration as this disconnection forces them to develop ad-hoc strategies and discourages them to revisit existing design knowledge.

Linking communication channels, such as, e-mail and IM, with project spaces will allow better integration of artifacts and related communication. When designers communicate through these channels, options should be provided to link the artifacts with the communications. Similarly, designers should be able to access relevant communication from their artifact spaces. Integrating plug-ins in the e-mail clients (e.g., Outlook) to create a bidirectional channel between the communication space and artifact space would allow designers to make this connection. Linking communications with artifacts would require some effort from the designer, but we believe that designers would be willing to invest this effort if it allows better retrievability.

5.2 Categorizing and Organizing Communication

Different types of communication influence the design process differently. While some communication allows the designer to select the best direction, others assist to improve the quality of artifacts. Some types of communication have value that extends well beyond the lifetime of a project, while others are only important during an ongoing design. Designers want to categorize and organize communication in a manner that facilitates effective access, retrieval and utilization. However, categorization of various types of communication introduces overhead in the design process, for which designers want to (and can) invest little effort.

Creating a vocabulary for capturing and categorizing different types of communication with related artifacts (e.g., requirement, user preference, design decision, and feedback) would allow effective access and retrieval. Incorporating default spaces (folders for client communication, feedback, rationale) for different categories of communication in the artifact space could reduce the burden of organization on the designer. Alternatively, tags could be utilized to associate different types of design communication with appropriate categories. These tags could be integrated in the e-mail and IM clients and designers could attach one or more tags to link the

communication (and relevant artifacts) to the appropriate category. Selecting a specific tag would provide access to all materials linked to that category.

5.3 Aggregate Communication from Multiple Channels

Designers engage in communication throughout the ideation process and obtain valuable information from various channels, such as face-to-face and phone conversations, e-mail, IM, online forums, etc. Lack of support for aggregation of communications obtained from various channels makes it difficult and time-demanding for designers to effectively utilize communication in the design process.

While it may not be possible to capture verbal conversations without any effort from the designer, communication that occurs through digital channels can be aggregated automatically. To facilitate lightweight capture of verbal communication, artifact spaces should incorporate free-form text entry spaces such as notes, which will allow designers to add communication outcomes with the artifacts. Allowing aggregation of communication obtained from various digital channels (e-mail, IM, forum) along with relevant design artifacts through lightweight linking could provide a solution to this problem. Add-ons in the e-mail and IM interfaces could be used to link pertinent communication to an artifact. Add-ons could be integrated in the online forums to allow designers to connect selected communications to the artifact space. Ideally, the artifact space should be connected to all communication channels and should provide support for linking communications received from various channels.

5.4 Support Finding Communication Partners by Integrating Interaction, Preference, and Availability Information

Designers almost always prefer to communicate with designers who can offer them valuable insight to improve the quality of their design. Expert designers often have access to suitable communication partners (other designers whom they deem fit), but it is difficult for novice and freelance designers to find appropriate communication partners. On the other hand, it can also become demanding for known experts in an area to accept communication (feedback) request from many designers.

A possible solution can be allowing designers to include communication preference and availability in their organization and/or online profile (in intra-networks, wikis). A visual history of projects they have worked on, clients they have worked for, their area of expertise and interest can be attached to their profile and also can be utilized for suggesting communication partner against a designers' query. For an online public setting, past communication can be analyzed including posts they have initiated and responded to. Designers' search for communication partners could be used along with their profile and interaction history to recommend a list of potential communication partners. Designers could select other designers from the suggested list or by analyzing the profiles that best meet their communication needs.

6 Conclusion

Communication surrounding artifacts is one of the key activities during the ideation process. Though designers utilize communication and design artifacts simultaneously

throughout the ideation process, existing research on design focus on these components separately; resulting in tools which make effective utilization of artifacts and related communication challenging. Better technology can assist to integrate artifacts with communication; however, in-depth understanding of communication practices in relation to artifacts is critical before building such systems.

This paper has contributed deeper understanding of communication practices in relation to artifacts during the ideation process. Our findings indicate communication outcomes greatly influence the generation, refinement, and selection of ideas during the early design stages. We also identify organizational and social parameters that influence designers' communication practice. We observed designers making genuine efforts, but struggling to capture communication outcomes with relevant artifacts, indicating a need for better design support tools. We offer guidelines that can be utilized to design new systems or to extend existing tools to better support the inter-connection of communication with relevant design artifacts.

References

1. Eckert, C., Stacey, M.: Sources of inspiration: a language of design. *Design Studies* 21(5), 523–538 (2000)
2. Sonnenwald, D.H.: Communication roles that support collaboration during the design process. *Design Studies* 17(3), 277–301 (1996)
3. Sutton, R.I., Hargadon, A.: Brainstorming Groups in Context: Effectiveness in a Product Design Firm. *Administrative Science Quarterly* (1996)
4. Bonnardel, N.: Creativity in Design Activities: The Role of Analogies in a Constrained Cognitive Environment. In: *Proc. Creativity & Cognition*, pp. 158–165 (1999)
5. Garner, S., McDonagh, D.: Problem Interpretation and Resolution via Visual Stimuli: The Use of 'Mood Boards' in Design Education. *Art & Design Education* 20(1) (2001)
6. Büscher, M., Kompast, M., Lainer, R., Wagner, I.: The Architect's Wunderkammer: Aesthetic Pleasure and Engagement in Electronic Spaces. *Digital Creativity* 10(1), 1–17 (1999)
7. Landay, J.A., Myers, B.A.: Interactive Sketching for the Early Stages of User Interface Design. In: *Proc. CHI*, pp. 43–50 (1995)
8. Regenbrecht, H., Haller, M., Hauber, J., Billinghamurst, M.: Carpeno: interfacing remote collaborative virtual environments with table-top interaction. *Virtual Reality* 10(2), 95–107 (2006)
9. Stewart, J., Bederson, B.B., Druin, A.: Single display groupware: A model for co-present collaboration. In: *Proc. CHI*, pp. 286–293 (1999)
10. Conklin, E.J., Burgess-Yakemovic, K.: A Process-Oriented Approach to Design Rationale. *Journal of HCI* 6, 357–391 (1996)
11. Lee, J.: SIBYL: a tool for managing group design rationale. In: *Proc. CSCW*, pp. 79–92 (1990)
12. Chiu, M.L.: An organizational view of design communication in design collaboration. *Design Studies* 23(2), 187–210 (2002)
13. Kraut, R.E., Fish, R.S., Root, R.W., Chalfonte, B.L.: *Informal communication in organizations: Form, function, and technology*. Morgan Kaufmann, CA (1993)
14. Nardi, B.A., Whittaker, S.: *The Place of Face-to-Face Communication in Distributed Work*. MIT Press, Cambridge (2002)

15. Eckert, C., Stacey, M.: Dimensions of Communication in Design. In: International Conference on Engineering Design (2001)
16. Stempfle, J., Badke-Schaub, P.: Thinking in design teams - an analysis of team communication. *Design Studies* 23(5), 473–496 (2002)
17. Herring, S.R., Chang, C.-C., Krantzler, J., Bailey, B.P.: Getting inspired!: understanding how and why examples are used in creative design practice. In: Proc. CHI, pp. 87–96 (2009)
18. Kolodner, J.: Improving Human Decision Making Through Case-based Decision Aiding. *AI Magazine*, 52–68 (1991)
19. Sharmin, M., Bailey, B.P., Coats, C., Hamilton, K.: Understanding knowledge management practices for early design activity and its implications for reuse. In: Proc. CHI, pp. 2367–2376 (2009)
20. Klemmer, S.R., Newman, M.W., Farrell, R., Bilezikjian, M., Landay, J.A.: The designers' outpost: a tangible interface for collaborative web site. In: Proc. UIST, pp. 1–10 (2001)
21. Klemmer, S.R., Thomsen, M., Phelps-Goodman, E., Lee, R., Landay, J.A.: Where Do Web Sites Come From? Capturing and Interacting With Design History. In: Proc. CHI, pp. 1–8 (2002)
22. Conklin, J., Begeman, M.L.: gIBIS: A Hypertext Tool for Team Design Deliberation. In: Proc. Hypertext, pp. 247–251 (1987)
23. Reeves, B., Shipman, F.: Supporting Communication between Designers with Artifact-Centered Evolving Information Spaces. In: Proc. CSCW, pp. 394–401 (1992)
24. Arias, E., Eden, H., Fisher, G.: Enhancing communication, facilitating shared understanding, and creating better artifacts by integrating physical and computational media for design. In: Proc. DIS, pp. 1–12 (1997)
25. Atwood, M.E., Burns, B., Gairing, D., Girgensohn, A., Lee, A., Turner, T., Alteras-Webb, S., Zimmermann, B.: Facilitating Communication in Software Development. In: Proc. DIS, pp. 65–73 (1995)
26. Lucero, A., Aliakseyeu, D., Martens, J-B.: Funky Wall: Presenting Mood Boards Using Gesture, Speech and Visuals. In: Proc. AVI, pp. 425–428 (2008)
27. Tang, J.C., Leifer, L.J.: A Framework for understanding the workspace activity of design teams. In: Proc. CSCW (1988)

Children's Interactions in an Asynchronous Video Mediated Communication Environment

Michail N. Giannakos¹, Konstantinos Chorianopoulos¹, Paul Johns²,
Kori Inkpen², and Honglu Du³

¹ Ionian University, Corfu, Greece

{mgiannak, choko}@ionio.gr

² Microsoft Research, Redmond, WA, USA

{paul.johns, kori}@microsoft.com

³ Penn State University, PA, USA

hzd106@ist.psu.edu

Abstract. Video-mediated communication (VMC) has become a feasible way to connect people in remote places for work and play. Nevertheless, little research has been done with regard to children and VMC. In this paper, we explore the behavior of a group of children, who exchanged video messages in an informal context. In particular, we have analyzed 386 videos over a period of 11 weeks, which were exchanged by 30 students of 4th and 5th grade from USA and Greece. We found that the number of views and the duration of a video message significantly depend on the gender of the viewer and creator. Most notably, girls created more messages, but boys viewed their own messages more. Finally, there are video messages with numerous views, which indicates that some videos have content qualities beyond the communication message itself. Overall, the practical implications of these findings indicate that the developers of asynchronous VMC should consider functionalities for preserving some of the video messages.

Keywords: Asynchronous, Video-Mediated Communication, Children, Thread, Gender.

1 Video Mediated Communication

Computer-mediated communication (CMC) includes a variety of electronic messages and audio-video systems. There is also increasing evidence that CMC mediums is replacing traditional forms of media and becoming a primary mode of communication in the workplace [1]. In this research, we are exploring the potential of CMC in the classroom, with a special focus on intercultural communication between distant places. Several researchers [2, 3] have claimed text-based asynchronous CMC to be incoherent for reasons such as the lack of simultaneous feedback, and the disrupted turn adjacency. However, video mediated communication (VMC) provides social context cues such as non-verbal signals (facial expressions, gestures); paraverbal cues (voice volume); and interpersonal cues (gender, physical appearance).

Various studies have shown that children are usually ineffective communicators because they have not mastered the necessary linguistic or cognitive competencies [4]. Bruner asserted that a language-based medium like email would be more complex for children than a medium that leveraged actions, bodily movement, or imagery. VMC is considered the most desirable to support nonverbal communication among children [5].

Few studies have reported gender differences in CMC as a main interest (independent variable). Hiltz and Johnson [6] found that females viewed CMC more favorably than males. Another notable research with an intra-organizational mail system, females believed e-mail to be easier to use, more effective and efficient than males [7]. Furthermore, Adrianson [8] mentions that females tended to produce more messages than males in the face-to-face communication, but in the CMC there was no significant difference. Another useful variable which characterizes discussion or/and conversation is the depth and the breadth. In our study we used the depth and the breadth of a thread (conversation) to explore the size of each conversation/discussion.

The current study focuses on using asynchronous video messages for children's communication. We evaluated a video-based asynchronous tool called Video Pal [9]. We were particularly interested in asynchronous video because it is an ideal tool to support communication between people from different parts of the world spanning many time-zones.

2 Methodology

The sample of participants in this study was comprised of 30 students. From the total of the participants, 25 pupils (12 boys and 13 girls) were from USA and 5 pupils (3 boys and 2 girls) were from Greece, as such 15 were boys (50%) and 15 (50%) girls.

One tradition in the USA partner school is that every year, each grade selects a country to study, and learns about that country's culture and lifestyles. In the final week of the fall semester, the children give presentations about what they have learned about that country to all the teachers, students and their parents. From the perspective of Greek pupils their key motivations was that it was regarded as a good opportunity for them to practice their English and it also enabled them to learn more about computers.

They employed the VideoPal user interface (figure 1), which has a main window and a message window. The main window allows users to quickly see which conversation threads are available, the properties of each thread (e.g., number of messages, number of unread messages in the thread), a visual presentation of one conversation, the new messages which are shown at bottom of the visualization panel, and the current users' profile photo. From this main window, users can create a new video or play an existing video message. The message window serves two purposes – to record a new message and play a received message. Users can play a received message, record a video message, and play a video preview, which is listed in the up right corner of this window.



Fig. 1. Video Pal main window (left), message view/creation windows (right)

The duration of the study was approximately eleven weeks. The study began on the 12th of November (2010) and finished on the 28th of February (2011) with breaks during Christmas (17 Dec- 11Jan) and 04-15 Feb. The following graph (figure 2) indicates the number of videos created each week and their length during the study. During the first week each country-group only sent videos internally, in order to get familiar with the user interface.

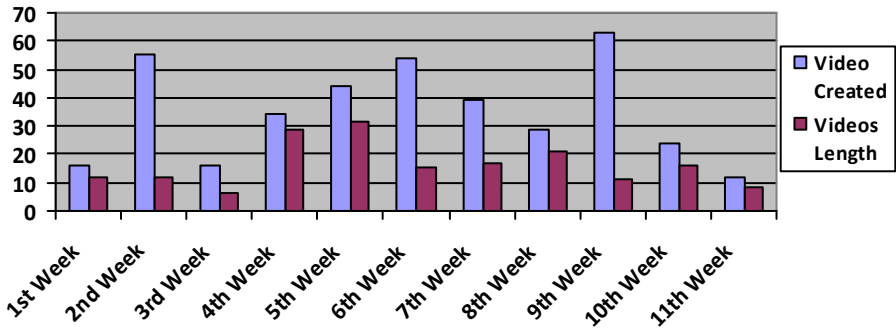


Fig. 2. The number of videos created each week and the average length of the videos each week

This paper presents an analysis of the computer log data. Descriptive data on all the video messages transmitted during the study were obtained through computer log files. In particular, we examined the following interactions: 1) number of author viewings, 2) number of recipient viewings, 3) video duration, 4) number of recipients and 5) distance of a video from the start of the thread (Video Distance). Authors’ viewings refer to the number of times the author of video replayed the video after sending it. Recipient viewings is the number of times a recipient viewed the video. Video duration is the length of each video in seconds. Number of recipients is the number of recipients per video message. The following table (table 1) outlines the descriptive statistics of the videos.

In general, key dimensions of communication are depth and breadth of the discussion-conversation [10, 11, 12]. As such, we used the video thread depth and the breadth in order to analyze the video discussions. Using VideoPal, the children could either reply to an existing video, or start a new thread/conversation. There were 171 threads created during this study. The mean number of videos in a thread was 2.26 with S.D.=1.11.

Table 1. Descriptive statistics of the videos (N=386)

	Country	No	%	Gender	No	%
Author	GR	84	21.8	Male	156	40.4
	USA	302	78.2	Female	230	59.6
Recipient(s)	GR	127	32.9	Male	110	28.5
	USA	139	36.0	Female	154	39.9
	Both	120	31.1	Both	122	31.6
	Mean (S.D.)					
Number of recipients	2.98 (3.26)					
Video Viewings	Author	0.78 (1.39)		Recipient(s)	6.26 (9.77)	
Video Duration (sec)	17.13 (15.19)					
Video Distance	1.93 (1.23)					

In our context, we defined the depth of a thread as the length of the longest chain in the thread. The breadth of the thread is the maximum number of replies to any video in the thread (figure 3).

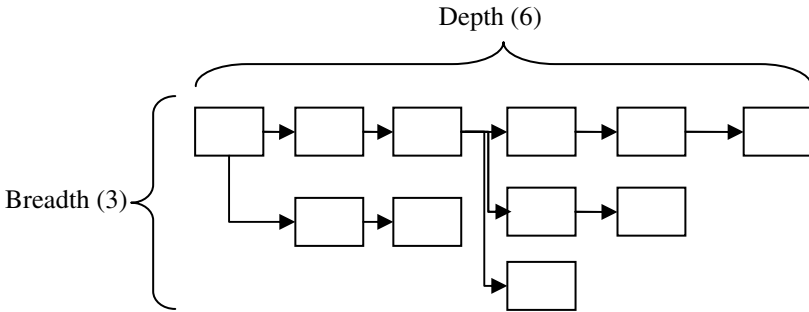


Fig. 3. An example of a thread depth and breadth

3 Research Findings

T-tests were conducted in order to investigate possible differences. T-test method was chosen, because it applies to the problem of estimating means [13]. Firstly we explored the impact of the gender of the pupil-author on number of viewings and the length of the video.

Using a t-test the following scores were deduced (table 2). The results showed that $t(212) = 3.51, p < 0.01$ and $t(293) = 4.04, p < 0.001$. This indicated that the number of viewings and the duration of the videos were significantly different between male and female authors. This leads us to the result that male authors view their videos more times and produce longer videos than female authors. Although females created shorter video messages, they also tended to create more video messages (70% more, as shown in table 1, although this difference was not statistically significant based on t-test result).

Table 2. Testing the impact of gender in authors viewings and video duration

	Male		Female		t-test			Results
	Mean	S.D	Mean	S.D	t	df	Sig.	
Author No of viewings	1.11	1.82	0.55	0.95	3.51	212	.001**	Significant Difference
Video duration	20.97	16.39	14.53	13.75	4.04	293	.000***	Significant Difference
Recipients no of viewings	3.72	4.18	3.23	4.34	.91	240	.361	No Significant Difference

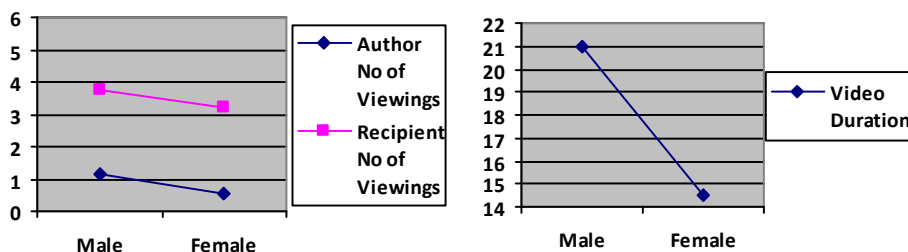
*** p<0.001

**p<0.01

*p<0.05

We also explored the impact of recipients' gender on the number of viewings. Using a t-test of two groups the following scores were deduced (table 2). The results showed that $t(240) = 0.91$, $p > 0.05$. This indicated that the number of times a recipient viewed a video was not statistically different for males and females.

As we noticed, the gender of the participants often had a significant effect on their interactions with the videos. The following figure (figure 4) shows the influence of gender and on pupils' viewings.

**Fig. 4.** The Influence of gender on user viewings and video duration (seconds)

We used Pearson's correlation coefficient between the features of the VMC, to quantify the strength of the relationship between the features. This analysis suggests that some of the features are related. For example, the number of recipients' viewings is strongly correlated with the number of recipients' and the distance of the video from the start of the thread (negatively). The duration of the video has strong negative correlation with the distance of the video from the start of the thread. Moreover, authors' number of viewings was significantly correlated with the number of recipients' viewings. All the correlations between the features are indicated in table 3.

Since the breadth and depth of the thread are two key dimensions of the conversation we chose to examine their behavior. However, an analysis of breadth and depth in small threads is not trustworthy, therefore, we made a depth-breadth analysis in threads, which were longer than 4 in depth (in total 8 threads). Figure 5 shows that as the breadth of the conversation grew, so did the depth.

Table 3. Pearson’s Correlation Coefficient between VMC features

VMC Features	Authors No of viewings	Recipients No of viewings	Video duration	Recipients Number	Dvsth
Authors No of viewings	1.000 (386)				
Recipients No of viewings	.125* (386)	1.000 (386)			
Video duration	.021 (386)	.063 (386)	1.000 (386)		
Recipients Number	.019 (319)	.561** (319)	-.017 (319)	1.000 (319)	
Dvsth	-.086 (386)	-.072* (386)	-.244** (386)	-.026 (319)	1.000 (386)

* Correlation is significant at the 0.05 level. ** Correlation is significant at the 0.01 level.

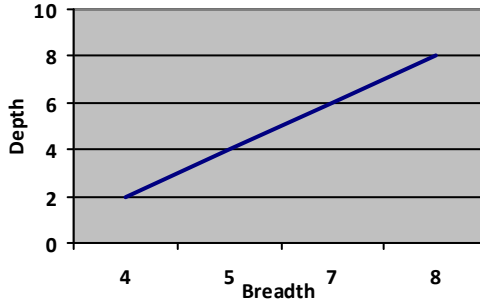


Fig. 5. Representation of breadth and depth behavior

4 Discussion and Ongoing Work

The above findings provide early evidence on VMC among children. In particular, the findings indicate that female video authors produce approximately 70% more videos than male authors. This finding is consistent with Hiltz and Johnson [6] results which found that females believed CMC to be easier to use, more effective and efficient than males. Moreover, Allen [7] states that females viewed CMC more favorably than males. However, Adrianson [8] found that females tended to produce more messages than males in face to face communication, but in CMC there was no significant difference. Therefore, this finding is not consistent with Tarasuik et al., [14] which states that for very young children video communication can have many of the same effects as a physical communication.

On the other hand, our findings indicate that males are viewing more of their produced videos than females; however, we found that females produce more videos. This may suggest that females prefer more active roles (produce) than males (view) in

VMC [15, 16]. In addition, females prefer to be more expressive at sending and decoding nonverbal messages and participate in more non-verbal communication behavior [17]. Interestingly, the findings indicate that males produce significant lengthier videos than females. This is consistent with Sussman and Tyson [18] which found that males produce longer postings in cyber-conversation.

The Pearson's correlation coefficients demonstrate that video duration has a strong negative correlation with the distance of the video from the start of the thread. This means that long duration videos are more likely at the beginning of a thread and short duration videos are most likely at the end of a thread. However this result may arise, at least in part, because video near the end of a thread indicate a communication flow, this communication flow led the participants in specific subjects and as a result in shorter videos. This is enforced by the fact that video based communication increased the feeling of "connectedness" between the participants [19].

Another interesting finding was that the number of authors' viewings was significantly related to the number of recipients' viewings and the number of recipients' viewings was significantly related to video depth. This result may arise from the nature of some videos; we can assume that the relation comes from the interest in some videos. However, we watched the most famous (high visibility of both authors and recipients site) videos and we realized that all these videos included useful personal information and the majority of them belong to a long depth videos. Interestingly, we observed these videos and we conclude that they were either introductory or descriptive for a participant. For instance, all these videos provided information such as, favorite foods, music, singer and hobby of the author; on the other hand videos with low visibility contain just questions. In brief, video messages with questions do not have repeat views, but video messages with statements can get a lot of attention, both by the creator and the recipients.

In an online context Parks and Floyd [11] and Parks and Roberts [12] found CMC users exhibit high levels of breadth and depth and form close relationships with their online partners. In our study we observed that the two key dimensions of each communication environment (depth and breadth) are following an absolute analogy behavior. This indicates that a video conversation keeps the analogies in their dimensions and they are not growing lop-sided.

As with any study, there are some limitations. For example, this pilot study was performed over a short time and the number of participants was quite small. Moreover, there was an asymmetry in the number of children between Greece and USA. We would also like to validate and extend the current findings with qualitative studies, such as longitudinal observations, or diary studies and investigate other challenges in cross-cultural VMC between children. Moreover, the follow up study would draw from a wider and more symmetric sample of children between the countries. In the light of an internal and external analysis of the groups we will identify which effects are caused from the culture of each group and which not. Finally, a deeper analysis of the video-user interactions using information such as timestamp of each view and video content analysis will provide a step ahead toward the understanding of children video mediated communication.

References

1. Tassabehji, R., Vakola, M.: Business email: The killer impact. *Communications of the ACM* 48(11), 64–70 (2005)
2. Herring, S.C.: Interactional coherence in CMC. *Journal of Computer-Mediated Communication* 4(4) (1999), <http://jcmc.indiana.edu/vol4/issue4/herring.html>
3. Marcoccia, M.: On-line polylogues: conversation structure and participation framework in Internet newsgroups. *Journal of Pragmatics* 36(1), 115–145 (2004)
4. Piaget, J.: *The language and thought of the child*. Brace & Company, Inc., NY (1926)
5. Ballagas, R., Kaye, J., Ames, M., Go, J., Raffle, H.: Family communication: phone conversations with children. In: *Interaction Design for Children 2009*, pp. 321–324 (2009)
6. Hiltz, R.S., Johnson, K.: User satisfaction with computer-mediated communication systems. *Management Science* 36, 739–751 (1990)
7. Allen, B.J.: Gender and computer-mediated communication. *Sex-Roles* 32(7/8), 557–563 (1995)
8. Adrianson, L.: Gender and computer-mediated communication: Group processes in problem solving. *Computers in Human Behavior* 17(1), 71–94 (2001)
9. Du, H., Inkpen, K.M., Tang, J., Roseway, A., Hoff, A., Johns, P., Czerwinski, M., Meyers, B., Chorianopoulos, K., Gross, T., Lungstrang, P.: Video Pal: An Asynchronous Video-Based Communication System to Connect Children from US and Greece. In: *Proc. of CSCW*. ACM Press, New York (2011)
10. Gay, G., Lentini, M.: Use of communication resources in a networked collaborative design environment. *Journal of Computer-Mediated Communication* 1(1) (1995)
11. Parks, M.R., Floyd, K.: Making friends in cyberspace. *Journal of Computer-Mediated Communication* 46, 80–97 (1996)
12. Parks, M.R., Roberts, L.D.: ‘Making MOOsic’: The development of personal relationships on-line and a comparison to their off-line counterparts. *Journal of Social and Personal Relationships* 15, 517–537 (1998)
13. Pallant, J.: T-tests. In: Pallant, J. (ed.) *SPSS. Survival manual*, 2nd edn., pp. 140–159. Open University Press, Berkshire (2005)
14. Tarasuik, J.C., Galligan, R., Kaufman, J.: Almost Being There: Video Communication with Young Children. *PLoS ONE* 6(2), e17129 (2011), doi:10.1371/journal.pone.0017129
15. Herring, S.: Gender differences in computer mediated communication: Bringing familiar baggage to the new frontier. Paper presented at the American Library Association Annual Convention, Miami, FL (June 1994)
16. Spender, D.: *Nattering on the net: Women, power and cyberspace*. Spinifex Press Ltd., North Melbourne (1995)
17. Briton, N.J., Hall, J.: Beliefs about female and male nonverbal communication. *Sex Roles* 23, 79–90 (1995)
18. Sussman, N., Tyson, D.: Sex and power: gender differences in computer mediated interactions. *Computers in Human Behavior* 16(4), 381–394 (2000)
19. Zuckerman, O., Maes, P.: Awareness system for children in distributed families. In: *Ext. Abst. of IDC*. ACM Press, New York (2004)

Effects of Automated Transcription Delay on Non-native Speakers' Comprehension in Real-Time Computer-Mediated Communication

Lin Yao¹, Ying-xin Pan², and Dan-ning Jiang²

¹ Institute of Psychology, Chinese Academy of Sciences, 10A Datun Road, Beijing, China

² IBM China Research Lab, Beijing, China

yaol@psych.ac.cn, {panyingx, jiangdn}@cn.ibm.com

Abstract. Real-time transcription generated by automated speech recognition (ASR) technologies with a reasonably high accuracy has been demonstrated to be valuable in facilitating non-native speakers' comprehension in real-time communication. Besides errors, time delay often exists due to technical problems in automated transcription as well. This study focuses on how the time delay of transcription impacts non-native speakers' comprehension performance and user experience. The experiment design simulated a one-way computer-mediated communication scenario, where comprehension performance and user experiences in 3 transcription conditions (no transcript; perfect transcripts with a 2-second delay; and transcripts with a 10% word-error-rate and a 2-second delay) were compared. The results showed that the participants can benefit from the transcription with a 2-second time delay, as their comprehension performance in this condition was improved compared with the no-transcript condition. However, the transcription presented with delay was found to have negative effects on user experience. In the final part of the paper, implications for further system development and design are discussed.

Keywords: Real-time transcription, Delay, Comprehension performance, User experience.

1 Introduction

Globalization is driving more and more people to communicate using their non-native language via audio/video conferences. However, as studies have indicated, understanding the speech of a second language often poses many difficulties [7]. Thus, non-native speakers frequently find it difficult to follow the conference and the collaboration tends to be ineffective.

Pan et al. [4] proposed an approach of using real-time speech transcription to improve non-native speakers' comprehension in long-distance communications, and further explored the possibility of using automatically generated transcription [5]. Their results indicated that when synchronized with the audio stream, the automated transcripts with a word error rate (WER) of 10% could significantly improve non-native speakers' comprehension, while transcripts with a WER greater than 20% would lead to no improvement in comprehension.

Unfortunately, in addition to recognition errors, time delay often exists in automated transcripts as well. The delay primarily results from the processing time the ASR system takes. Even using the most advanced speech recognition technology, the ASR processing can still be behind the audio stream for complex recognition tasks or low-quality signals. For a distributed ASR service hosting system like the one described in [2], the ASR processing will slow down when many concurrent users request the ASR service at the same time. The network delay for data transmission between the speech recognition client, server, and text showing interface also contributes to the overall delay. This study will investigate how time delay affects the usefulness of automated transcription in improving non-native speakers' comprehension in real-time communication.

Most previous studies on the effects of transcription delay have focused on helping people with hearing impairment to better understand audio or video contents [1,3,8]. Burnham et al. [1] and Maruyama et al. [3] examined hearing-impaired people's enjoyment and intelligibility of TV when captions delayed from 0 second to 4 seconds. They reported that both enjoyment and intelligibility diminished when the delay existed, and the permissible limit for delay varied from 1.63 to 4.84 seconds depending on the degree of hearing impairment and caption formats. Zekveld et al. [8] measured the benefits of transcription with speech reception threshold (SRT), and reported that delaying the transcription with 2, 4, or 6 seconds reduced the benefits of transcription by approximately 1 to 2 dB of SRT.

While these findings provided valuable insights into how time delay affects the usefulness of transcription, none of them touched upon using automated transcription to improve non-native speakers' comprehension. Furthermore, as non-native speakers need extra cognitive efforts to process the transcripts in a second language as one additional source of information, the delay could distract attention and result in little value of transcripts.

In this paper, we report an experiment that examines the effects of time delay of transcription on non-native speakers' comprehension performance and user experience, in which two research questions are addressed:

- Does automated transcription still help non-native speakers' comprehension in computer-mediated communication when a reasonable level of time delay exists in the transcripts?
- How does the time delay of transcription affect non-native speakers' user experience?

2 Method

2.1 Preliminary Study

Before the main experiment, we did a preliminary study to find out a time delay level worth being studied more thoroughly. We started from 2 seconds and 4 seconds, which, according to previous studies [1,3,8], might be a critical level of time delay for the transcription to be usefully and acceptable. 12 Chinese participants were divided into 2 groups and were asked to watch 6 English clips in 3 conditions: no transcript

was displayed (NT), transcripts with no delay (PT) and delayed transcripts (DT). Transcription delay of the two groups was set as 2 seconds and 4 seconds respectively. After each clip was played, the participants were asked to answer 5 comprehension questions to evaluate how well they understood the materials.

The results showed that when delay was 4 seconds, the transcripts did not help the comprehension. The comprehension score of using the delayed transcripts was even worse than that when no transcript was displayed. All participants reported that they felt really frustrated by the delay and preferred to just ignoring the transcripts. In contrast, when the transcription delay was 2 seconds, the comprehension performance was improved compared to the no transcript condition, though some of the participants still reported that the delayed transcripts were somewhat distracting. Thus, in the formal experiment, we will use 2 seconds delay to confirm the usefulness of delayed transcripts.

2.2 Experiment Setup

Similar to [4,5], we designed a one-way computer-mediated communication (CMC) scenario, in which native English speakers talked in English via an audio and video channel, and native Chinese "listeners" (the participants) tried to understand what was spoken. Though communication in this study was dominated by one or a few main speakers and others just listen, the findings or conclusions were believed to serve as a useful reference for future research on more interactive scenarios.

Figure 1 showed an example of the interface developed for the experiment. Transcripts were displayed in a streaming mode, appearing letter by letter from bottom left to right. This display mode is necessary in real-time scenarios as the speakers' words cannot be foreseen before being spoken.



Fig. 1. An interface example of the experiment design

2.3 Experiment Design

The formal experiment was designed as a within-subject study in which participants were exposed to three different Transcription Conditions:

- NT: No transcript was displayed (the baseline case).
- DT: Transcripts with 2 seconds' delay were displayed. No error was included in the transcripts.
- D-ET: Transcripts with 2 seconds' delay and 10% WER were displayed. This condition was to examine if automated transcripts with errors could help when they were not synchronized with the audio stream. The 10% WER level was selected because it was the best accuracy that could be achieved in practice (with high-quality signal, in-domain language model, and native accent) and thus might serve as the benchmark for the most tolerable level of time delay.

2.4 Participants

Thirty highly motivated university or graduate school students from various disciplines were recruited as participants. They were non-English major native Chinese speakers and had passed CET-6 (College English Test Band 6), a national English test which is mandatory for all Chinese students if they are to get a master's degree. A curious observation, however, is that though CET-6 indicates a relatively high level of English proficiency of Chinese students, there is no guarantee that those who have passed the test can understand spoken English conversations well. The participants were of a mixed gender (16 females and 14 males), and with an average age of 23.8 years ($SD=4.5$, range from 20-28).

2.5 Materials and Task

Six English video clips were created, 2 for each transcription condition (NT, DT and D-ET). The clips were 3.5-minute-long on average, and covered a broad range of general topics (e.g. advertising, environmental protection, obesity, etc.). 3 clips were dialogues cut from an English TV show, and the other 3 were lectures recorded with invited foreigners as speakers. 5 comprehension questions were designed for each clip, including both short-answer questions and multiple-choice questions. All the materials had been validated in our previous research and their difficulty level was appropriate for the Chinese participants [5].

The whole experiment was computer-based. A Latin square design was implemented to counterbalance order effects. Each participant was asked to watch the 6 clips. After each clip was played, the screen turned to the question-answer page immediately and no transcript could be seen any more. After finishing the comprehension test in each Transcription Condition, the participants were asked to complete a follow-up questionnaire on user satisfaction and cognitive load. The whole procedure of the experiment took about 60 minutes on average.

2.6 Measurements

Comprehension Performance. Performance was measured by response accuracy, that is, how many comprehension questions were answered correctly. A perfect score in each condition was 10 (5 questions*2 clips).

User experience. User experience was assessed by user satisfaction and user cognitive load.

User satisfaction. Participants were required to respond to three satisfaction evaluation questions on a 5-point Likert scale. The three questions were: (1) Usefulness: "I think transcription is helpful for my understanding." (2) Importance: "I think transcription is important for my understanding." (3) Preference: "I would love to have transcription next time."

Cognitive Load. Investigated how well human resources could be employed in task completion or problem solving. Three indicators were used:

- *Perception of task difficulty.* The participants assessed the difficulty of answering the questions by indicating their agreement with the following statements on a 5-point Likert scale: "It was difficult for me to correctly answer the comprehension questions" and "I fully understood what the clips talked about."
- *Perception of concentration difficulty* measured how well one can focus their cognitive resources on the task by asking the participants to respond to the following statement: "It was difficult for me to concentrate my attention simultaneously on the information from all sources (e.g., audio, video and transcription) ."
- *Perception of understanding interference.* The participants assessed how the time delay of the transcription might interfere with their understanding by indicating their agreement with the following statements on a 5-point Likert scale: "The time delay of transcripts distracted my attention" and "The time delay of transcripts hindered my understanding of video clips".

3 Results

All the data were submitted to SPSS 15.0 for analysis.

3.1 Comprehension Performance

The comprehension performance scores in different conditions were shown in Figure 2. A repeated measures ANOVA was used to analyze the data. The results showed that *Transcription Condition* had a significant main effect on performance, $F(2, 58) = 7.27, p < .01$, indicating that comprehension was indeed influenced by the transcription condition.

To further explore the difference between the comprehension performance in NT, DT, and D-ET, multiple comparisons were performed. The comprehension performance in DT, D-ET was found to be significantly better than that in NT (both $p < .01$). Performance in DT was a little better than that in D-ET (5.34 vs. 5.07), but the difference did not reach a significant level ($p > .05$). These results suggested the usefulness of automated transcription when the time delay is less than 2 seconds.

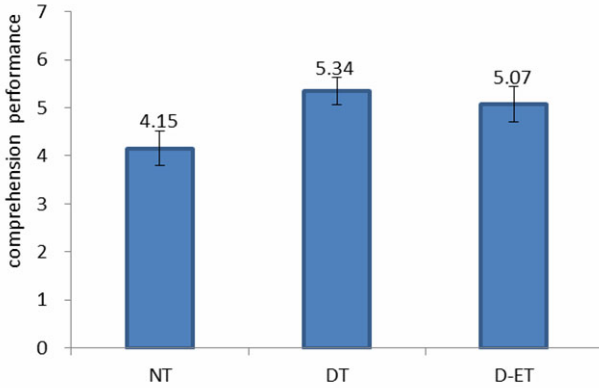


Fig. 2. Comprehension performance in NT, DT, and D-ET conditions. Error bar stands for one standard error.

3.2 User Satisfaction

The user-reported satisfaction scores for DT and D-ET condition were shown in Figure 3. The participants confirmed the usefulness of the delayed transcripts (the usefulness score in DT and D-ET condition was 4.21 and 4.03 respectively), while the importance and preference scores were nearly neutral. The results indicated that the participants did not feel very pleasant with the delayed transcripts though they were regarded as being useful

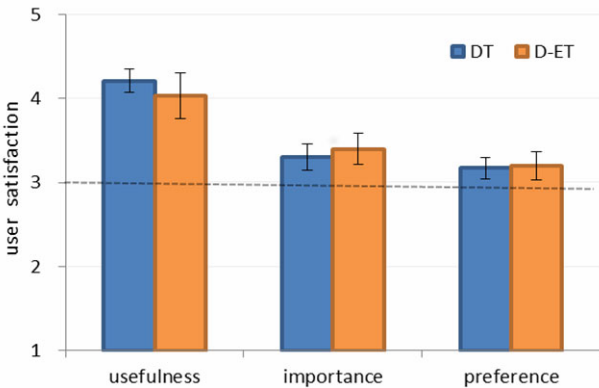


Fig. 3. User satisfaction in the DT and D-ET conditions

3.3 Cognitive Load

Cognitive load scores were presented in Table 1 in three dimensions. For task difficulty, it was found that the delayed transcripts did not increase the perceived

difficulty of the comprehension test ($p > .05$, repeated measures ANOVA). With regard to the perception of concentration difficulty and understanding interference, the results suggested a negative impact in general. The majority of the participants (66.7% in DT, 70% in D-ET,) agreed or strongly agreed that it was difficult to concentrate their attention simultaneously on the information from all sources (the means were 3.97 and 4.03 respectively). In addition, over half of the participants (56.7% in DT, 63.3% in D-ET) agreed or strongly agreed that the time delay of the transcripts would interfere with their understanding (the means were 3.80 and 3.87 respectively).

Table 1. Users' cognitive load in the NT, DT, and D-ET conditions (data were presented as means on a 5-point scale)

Cognitive load	NT	DT	D-ET
Task difficulty	2.70	2.47	2.55
Concentration difficulty	N/A	3.97	4.03
Understanding inference	N/A	3.80	3.87

4 Discussions, Conclusions, and Future Work

In this paper, we investigated how time delay in automated transcription produced by a speech recognition system affects non-native speakers' comprehension and user experience. The results demonstrated the value of delayed transcription in improving non-native speakers' comprehension in one-way communication scenario. When the time delay was 2 seconds, the participants' comprehension performance was significantly improved with the aid of the transcripts, and the users' self-reported satisfaction also confirmed the usefulness of the transcripts. But the users' self-reported measures still showed some negative effects of time delay, e.g. the increase of concentration load, the distraction time delay has causes, and its interference with the understanding of audio information.

It seems somewhat surprising that while the non-native speakers' comprehension performance was factually improved by using the transcripts, they still reported some negative user feelings. This can be explained from several aspects. First, the task being simulating the passive one-way communication, instantaneous response on the part of the users was not required. Thus, despite the time delay, the appearance of the transcripts provided a chance for gist extraction and therefore improves the comprehension [6]. Second, the users had to pay more attention and work harder when there was time delay in the transcripts. The increased attention would result in better comprehension. But paying more attention and working harder would be more stressful and decrease the satisfaction. Third, the output delay was not only obvious enough to be perceived by users but also rendered this type of transcription very much different from ordinary types of transcription (e.g. DVD captions) with which users are already quite familiar, hence they are inclined to make a negative assessment.

In summary, this study demonstrates that automated transcription in a good accuracy and with a reasonable level of delay (≤ 2 seconds) can significantly improve

non-native speakers' comprehension, though user experience evaluations are not all positive. The paper implies that a certain level of transcript delay which temporally happens at system busy time (e.g. caused by large number of concurrent users) can be acceptable. But since time delay would result in negative user experience, the system should ensure that important conferences can get sufficient computation resources and high quality network connection to avoid the delay.

Future work will investigate the effects of time delay in more interactive scenarios involved in remote collaborations. In addition, finer levels of word error rate in automatically generated transcripts combined with delays should be studied, as WER in automated transcription could change within a broad range from 10% to over 30%.

References

1. Burnham, D., Robert-Ribes, J., Ellison, R.: Why captions have to be on time. In: Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP 1998), pp. 153–156 (1998)
2. Jiang, D., Pan, Y., Liu, W., Qin, Y., Picheny, M., Luther, P.: Real-time Speech Transcription Service to Improve Non-native Speaker's Listening Comprehension. In: The 25th Annual International Technology & Persons with Disabilities Conference (2009)
3. Maruyama, I., Abe Y., Sawamura E., et al.: Cognitive Experiments on Timing Lag for Superimposing Closed Captions. In: Proceedings of the Sixth European Conference on Speech Communication and Technology, pp. 575–578 (1999)
4. Pan, Y., Jiang, D., Picheny, M., Qin, Y.: Effects of real-time transcription on non-native speaker's comprehension in computer-mediated communications. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, pp. 2353–2356. ACM Press, New York (2009)
5. Pan, Y., Jiang, D., Yao, L., Picheny, M., Qin, Y.: Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, pp. 1725–1734. ACM Press, New York (2010)
6. Tucker, S., Kyprianou, N., Whittaker, S.: Time-Compressing Speech: ASR Transcripts Are an Effective Way to Support Gist Extraction. In: Proceedings of the 5th Joint Workshop on Machine Learning and Multimodal Interaction, pp. 226–235 (2008)
7. Tyler, M.D.: The Effect of Background Knowledge on First and Second Language Comprehension Difficulty. In: Proceedings of the 5th International Conference on Spoken Language Processing (1998)
8. Zekveld, A.A., Kramer, S.E., Kessens, J.M., Vlaming, M.S., Houtgast, T.: The influence of age, hearing, and working memory on the speech comprehension benefit derived from an automatic speech recognition system. *Ear and Hearing* 30, 262–272 (2009)

Redundancy and Collaboration in Wikibooks

Ilaria Liccardi^{2,1}, Olivier Chapuis^{1,2}, Ching-Man Au Yeung³, and Wendy Mackay^{2,1}

¹Univ. Paris-Sud & CNRS, Orsay, France

²INRIA, Orsay, France

³NTT Communication Science Laboratories, Kyoto, Japan

Abstract. This paper investigates how *Wikibooks* authors collaborate to create high-quality books. We combined Information Retrieval and statistical techniques to examine the complete multi-year lifecycle of over 50 high-quality Wikibooks. We found that: 1. The presence of redundant material is negatively correlated with collaboration mechanisms; 2. For most books, over 50% of the content is written by a small core of authors; and 3. Use of collaborative tools (predicted pages and talk pages) is significantly correlated with patterns of redundancy. *Non-redundant* books are well-planned from the beginning and require fewer talk pages to reach high-quality status. *Initially redundant* books begin with high redundancy, which drops as soon as authors use coordination tools to restructure the content. *Suddenly redundant* books display sudden bursts of redundancy that must be resolved, requiring significantly more discussion to reach high-quality status. These findings suggest that providing core authors with effective tools for visualizing and removing redundant material may increase writing speed and improve the book's ultimate quality.

Keywords: Collaborative writing, text redundancy, coordination mechanisms.

1 Introduction

The advent of the World Wide Web and wiki-based collaboration technologies has made it possible for a new form of mass collaboration in which groups of strangers work together on a common topic. These on-line, volunteer-based projects have produced major new resources. One of the most successful examples is the Linux kernel, which was developed by a large number of unpaid contributors [27]. More recently, wiki technologies have been introduced to facilitate the creation and editing of inter-linked pages, e.g. the Wikipedia encyclopedia and Wikibooks. Unlike collaborative writing within corporate environments, wiki technologies permit large numbers of strangers from around the world to work together on a shared topic.

This type of collaborative writing has inherent advantages and disadvantages. Each project may benefit from a wealth of expertise and knowledge but faces enormous coordination and communication challenges in order to produce a coherent final result. Manuscripts typically evolve during the writing process and discussions and disagreements inevitably occur, due to differences in knowledge, experience and points of view. Groups of co-authors manage their work differently, which affects writing speed, manuscript structure, the validity of the arguments and the perceived quality of the text itself.

In traditional collaborations, contributors are linked by professional ties. Thagard [31] identifies three types of collaboration that reflect the backgrounds and roles of the authors: dominant relationships, e.g. employer/employee or teacher/apprentice, peer relationships, e.g., among researchers with similar backgrounds, knowledge, and skills, and peer-different, e.g., among researchers from different disciplines who share similar goals. However, even in corporate environments, studies of informal collaboration [23, 24] show that motivated individuals sometimes voluntarily take on coordination tasks to support their colleagues.

Unlike projects that occur within a corporate hierarchy, open-source collaborations are often driven by what Lerner calls "hobbyists" [19] who do not have clearly defined roles with respect to each other. Thagard's classifications are particularly difficult to define when authors do not interact directly with each other. This raises the question of how large-scale, volunteer-driven writing projects can coordinate the efforts of large numbers of users and still produce high-quality results.

Researchers in CSCW have begun studying large-scale writing projects, hoping to gain insights into how best to design tools to facilitate collaboration. To date, Wikipedia is most well-studied [14, 16, 22]. Researchers have examined communication patterns, conflict resolution and authorship and has used some of these findings to design tools to support collaboration [3]. They disagree about the benefits of including very large numbers of participants. On the one hand, new authors offer the potential for gaining additional expertise and novel perspectives, thus increasing the value of the result [4, 11]. However, adding authors may also reach a point of diminishing returns, with a trade-off between the benefits of additional resources and the costs of increased coordination [10].

Brook's Law [1] famously argues that additional coordination costs can easily overwhelm any benefits from added personnel: "Adding manpower to a late software project makes it later". For this reason, corporate managers and book editors often choose to limit group size to increase productivity [29]. In stark contrast is Raymond et al.'s [27] claim that involving as many authors as possible improves open-source software projects. They suggest that it is important to "delegate everything you can, be open to the point of promiscuity". Kittur and Kraut [14] refine this claim, showing that coordination is essential for harnessing the wisdom of the crowds. They studied the relationship between the number of authors and the use of appropriate coordination tools on the quality of Wikipedia content. Essentially, articles with many authors are better, but only if the authors can effectively coordinate their activities. Of course, this also depends upon the type of work. Stewart [30] shows that larger teams generally perform better when they engage in low-coordination as opposed to high-coordination work.

We are interested in a less well-known, but no less interesting, collaborative writing system, called Wikibooks. The primary goal of Wikibooks is to provide free, printable textbooks that can be used in the classroom. Although most are educational [28, 34], they also cover a wide range of other topics, including sports, religion, interpersonal relationships and even a guide to Harry Potter novels.

Wikibooks is based on the same MediaWiki software and open editing policy as Wikipedia. However, Wikibooks co-authors face a greater challenge than Wikipedia contributors, because they must coordinate their activities on a much greater scale,

over much longer periods of time. The longest Wikipedia article is around 50,000 words whereas over 100 Wikibooks exceed this.

In order to study how large groups of strangers coordinate their activities over long periods of time, we examined the first complete set of Wikibooks logs, dating from its creation in 2003 until 2009. Our goal was to identify the key coordination and communication mechanisms that affect quality.

Our first challenge was to find appropriate measures of quality. Studies of Wikipedia have identified a diverse set of possible metrics, including: number of edits and unique editors [22], factual accuracy [8] (but disputed by [6]), credibility [2], revert times [33], and the formality of language [5]. In each case, researchers applied these metrics to small samples of Wikipedia articles and, sometimes, to equivalent articles in traditional encyclopedias as an independent standard of quality.

However Wikibooks involve even larger scale collaborative efforts, which renders some of these measures infeasible. We decided to focus on the 59 ‘featured’ Wikibooks, which had been designated of the highest quality, and use both Information Retrieval and statistical techniques to analyze the historical data. We focused on four key measures: level of redundancy in the text, authorship patterns, use of *predicted pages* and use of *talk pages*.

This paper begins with a description of the Wikibooks corpus, followed by the study design and our analysis methods. We then describe and discuss the results of three successive analyses: 1. Redundancy, 2. Co-authorship and 3. Collaboration lifecycle. We then provide an in-depth look at a few specific books and conclude with implications for design and directions for future research.

2 Wikibooks

We analyzed data obtained from the English version of Wikibooks¹. Since its beginning in 2003, the site has expanded to include over 2000 books, the content of which is contributed entirely by volunteers.

2.1 Corpus

The Wikibooks site includes books in all stages of development. The complete history of each page on Wikibooks is stored in a database on the website and it is possible to access every past revision of a page through the web interface. The Wikimedia Foundation provides downloadable versions of the database including the page history. Our data set comes from a database dump of the Wikibooks Website on 15 May 2009² and contains 2,039 books. Because we were interested in understanding the factors that are correlated with high-quality books, we focused our analysis on the 59 Featured books.

Of these 59, we removed eight books: Five were from the Wikijunior series, which are designed to be age-appropriate non-fiction books for children from birth to age 12.

¹ <http://en.wikibooks.org/>

² <http://download.wikimedia.org/enwikibooks/>

These books have an atypical structure, with very little text and many images. The co-authors discuss the content thoroughly before beginning to write to ensure that it is suitable for the intended audience. We also removed *FHSST PHYSICS* since this book's content was based on an existing text written outside of the Wikibooks environment. Finally, we removed *ADVENTIST YOUTH HONORS ANSWER BOOK* and *SOCIAL AND CULTURAL FOUNDATIONS OF AMERICAN EDUCATION* since they were structured more like catalogs or reference sources rather than actual books. We analyzed several statistics from these books to understand the level of participation in each book from its beginning to the time that the book was designated as a 'featured book'.

2.2 Featured Books as a Measure of Quality

Wikibooks does not provide a quantitative measure for book quality, so we chose an empirical, but qualitative measure, 'featured book' status. The Wikibooks community identifies a small number of books that they judge to be of high quality, based on a set of pre-defined criteria. These criteria are imprecise (probably deliberately so), but do offer guidelines as to what constitutes a good book³. Criteria include, for example, the clarity of the text, the structure and completeness of the book, and whether or not the book conforms to Wikibooks policies.

Any Wikibooks reader can nominate a book for 'featured book' status. Once a number of users have so nominated a book, an administrator reviews the strength of the arguments with respect to the above criteria. If it passes this test, the book is voted on through a democratic process. Successful books are then added to the list of 'featured books'⁴. Only a small subset of Wikibooks is nominated and they represent about 3% of the books available on the site. The guidelines are stringent and these books must maintain their quality in order to maintain featured status. Because volunteers may edit books at any time, the Wikibooks community actively monitors the quality of these books and removes them if necessary.

Previous studies of Wikipedia [14, 15, 16] used a similar measure of quality, based on whether an article was explicitly chosen to be featured by Wikipedia administrators. In other studies, Wikipedia users were asked to read an article and rate its quality [13]. However, as mentioned earlier, since Wikibooks are much longer than encyclopedia entries, the latter strategy is not practical for our purposes.

Featured and Non-featured. Because we were interested in understanding the collaborative process involved in writing *successful* books, we focused solely on featured books. We calculated measures of redundancy from the inception of each book through to its completion, as indicated by its election to 'featured book' status. Although these books comprise only a small percentage of the total number books, they are much longer on average and account for almost half of the total size of the Wikibooks database. We recognize that some non-featured books are also close to completion, but since we have no objective, independent criteria by which to measure the quality of such books, we do not include them.

³ http://en.wikibooks.org/wiki/Wikibooks:Good_books

⁴ http://en.wikibooks.org/wiki/Wikibooks:Featured_books

3 Analysis 1: Redundancy Patterns

One possible measure of how authors coordinate their activities is *redundancy*. We hypothesized that redundant text can act as a proxy for communication breakdowns and a lack of coordination among participants. We also expected that redundant text will be highly correlated with other negative indicators, such as the presence of cleanup and maintenance tags. We thus measured redundancy within individual books, looking for patterns of how it changed over time. At regular intervals during the evolution of each book, we took a snapshot and analyzed the full text of the manuscript at that point, and counted the number of redundant paragraphs.

Note that redundancy can only be interpreted within the context of the particular book. Because different Wikibooks are structured differently and cover completely different subject areas, it does not make sense to compare levels of redundancy across books: some books may require some redundancy whereas others do not. Thus, our classification is not based on absolute redundancy values, but rather on an analysis of how redundancy curves change over time, within the context of each individual book.

Research Questions. We are particularly interested in:

- How does the level of redundancy change over time?
- Do different books exhibit different patterns of redundancy from their beginning to completion?

3.1 Measuring Redundancy via Semantic Similarity

We base our measure of redundancy on argument *repetition*, i.e. the amount of similarity among arguments (sentence, recurring words etc.), at the level of paragraphs, sections and chapters, at different points in the writing process. This provides us with an objective, quantitative measure of the effectiveness of the different communication and collaboration mechanisms. Although redundancy is sometimes used for automatic document summarization, in which redundancy in the summary or abstract is used as an indicator of poor quality, we are unaware of other studies that use redundancy as a measure of coordination.

Our strategy is to quantify the level of redundancy by determining the semantic similarity between two sentences. Although this is challenging even with modern natural language processing techniques, a combination of techniques has proven to be effective, e.g. [9, 20], and offers an approximation for the amount of similarity and thus redundancy between two sentences.

Words are the most basic unit of measure for identifying similarity between texts. Techniques for measuring similarity are based on string similarity [12], thesauruses [26] or corpus statistics [32]. In information retrieval, the ‘bag of words’ approach is commonly used to measure similarity between documents. A document is usually characterized by a term vector of length n , the elements of which indicate whether a term is present in the document and its relative importance. Terms are usually weighted using the TF-IDF (term frequency-inverse document frequency) scheme [17]. The cosine similarity measure is then used to compare the two vectors.

Other methods have also been proposed that exploit word co-occurrence information instead of using exact word matching. For example, latent semantic analysis [18]

can be used to measure similarity between texts by computing higher-order word relations based on dimensionality reduction. Some approaches combine different techniques. For example, Li et al. [21] propose a method that combines Word Net based word similarity, corpus statistics and word order similarity. Islam and Inkpen [12] propose a similar approach based on substring matching of words, point-wise mutual information similarity, and word order similarity.

Redundancy Measure. We chose to use cosine similarity between term vectors constructed by using the TF-IDF weighting scheme [17] to measure redundancy between paragraphs. We decided to use cosine similarity to measure pairwise similarities in each snapshot of each Wikibook.

In formal notation, let T be the set of terms that appear in a Wikibook B . We employ standard Information Retrieval preprocessing methods, such as stop-word removal and stemming [35] to produce the set P of paragraphs. Each paragraph $p \in P$ is characterized by a term vector:

$$v_p = (w_{p,1}, w_{p,2}, \dots, w_{p,|T|})$$

where $w_{p,i}$ is the weight of term $t_i \in T$ given by the standard TF-IDF weighting scheme.

The similarity between two paragraphs p and q can then be calculated by using the cosine similarity measure, given by:

$$sim(v_p, v_q) = \frac{v_p \cdot v_q}{\|v_p\| \times \|v_q\|}$$

The cosine similarity tells us only how similar two paragraphs are with respect to the terms they contain, but does not tell us how much redundancy is observed in the book. Here, we define redundancy as the proportion of pairs of paragraphs that attain a similarity value higher than a threshold α :

$$redundancy(B) = \frac{\sum_{p,q \in P, p \neq q} \delta(p,q)}{2 \times C_2^{|T|}} \text{ where } \delta(p,q) = \begin{cases} 1 & \text{if } sim(v_p, v_q) \geq \alpha \\ 0 & \text{otherwise} \end{cases}$$

and since there are $C_2^{|T|}$ pairs of (p,q) and the similarity function is symmetric, the term $2 \times C_2^{|T|}$ is used to normalize the redundancy score.

Measure Refinement and Validation. While a high level of similarity between two paragraphs does not always imply the existence of redundant information, in practice, cosine similarity can be used to approximate redundancy.

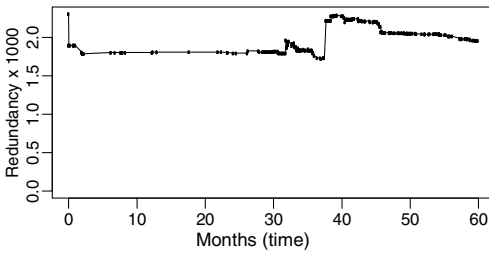
We validated this measure by asking six volunteers to read and classify the appearance of redundancy in a randomly selected sample of paragraphs. We took a sample of 603 paragraphs (201 for each category). We then randomized the paragraphs into 9 sets, each containing 67 paragraphs for each set, making sure that at least two volunteers examined each text to ensure reliability. We found that a threshold of 0.5 was able to correctly identify redundant paragraphs. In fact, we found that,

in 90% of the cases, the text identified by the redundancy algorithm was indeed designated as redundant by the volunteers. For the remaining 10%, the text was identified as redundant because the same quotations were repeated across paragraphs.

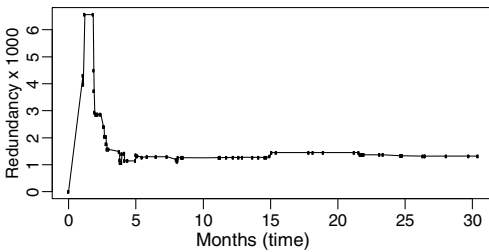
3.2 Results

We measured changes in redundancy from when the book was started until its completion, i.e. when it attained ‘featured’ status. (Fig. 1 shows the patterns of redundancy levels over time for three typical books.) We found that none of the 51 featured books contained redundant text when they were completed. However, we did find that books fell into one of three main categories with respect to how redundancy levels changed over the lifetime of the book:

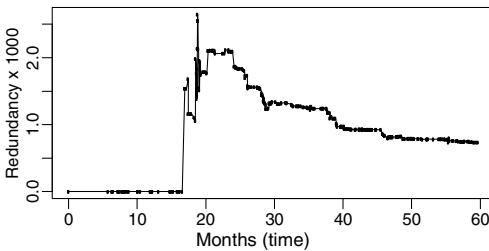
- *Non-Redundant* books (14/51) include negligible amounts of redundancy throughout the book. (Fig. 1.a: `NON-PROGRAMMER’S TUTORIAL FOR PYTHON`).
- *Initially Redundant* books (23/51) begin with a great deal of redundancy, followed by either a slow, steady decline or a rapid drop in redundant material. (Fig. 1.b: `THIS QUANTUM WORLD`).



a. *Non-Redundant:*
`NON-PROGRAMMER’S TUTORIAL FOR PYTHON`



b. *Initially Redundant:*
`THIS QUANTUM WORLD`



c. *Suddenly Redundant:*
`C# PROGRAMMING`

Fig. 1. Evolution of redundant text over time (in months). Values are not absolute.

- *Suddenly Redundant* books (14/51) begin with low levels of redundancy, followed by sharp increases at different points during the development process. The level of redundancy then decreases gradually over time. (Fig. 1.c: C# PROGRAMMING).

We next examined how the above categories affect coordination during the development process. In particular, we examined their correlation with the introduction of redundant text and the perceived quality of each book.

4 Analysis 2: Co-authorship

With ordinary edited books, a few co-authors divide the work and share top billing. However, in on-line wiki environments, “open-editing” means that potentially thousands of authors may contribute text, with contributions ranging from major sections to just a few words. We are interested in understanding how large numbers of authors who are strangers collaborate to create a ‘featured boo’. We thus examine redundancy with respect to other measures, including number of authors, duration of the writing process and length of the book.

Research Questions. We are particularly interested in:

- What is the distribution of authors and their edits in featured books?
- Do the number of authors and the distribution of their edits affect observed redundancy patterns?

4.1 Measures

For each book, we collected metadata about its edits: the authors, the number of revisions, and the times when the book was edited. We accessed all versions of the pages within the book, including associated talk pages. We also used the difference between the timestamp on the first revision and the date when the book was classified as featured to calculate its age. We used the three redundancy patterns `RedundancyPatterns = Initially-Redundant, Suddenly-Redundant, Non-Redundant` as a factor to analyze these measures.

4.2 Results

We used linear models to examine the possible correlations among continuous measures. Below, we report the p-values for the correlation slopes and the adjusted r^2 that measures the quality of fit. We also conducted one-way ANOVA analyses to examine the effect of the `RedundancyPatterns` on various measures. In turn, these analyses help us to understand the effect of these measures on redundancy patterns.

Authors’ Distribution. We can identify different contribution patterns within the histories of different books. The number of contributing authors is highly variable, ranging from 8 to 2631, with a mean of 272 ± 405 , a median of 159, a 10% quartile of 34 and a 90% quartile of 614 contributors (see Fig. 2.a).

One common pattern involves a small core of lead authors who write the bulk of the book, aided by a large number of supporting authors who may change only a few words. Panciera et al. [25] found a similar pattern among Wikipedia authors. If we rank authors by the number of their edits, the resulting frequency distribution resembles a zipf⁵ distribution [36], as in Fig. 2.b.

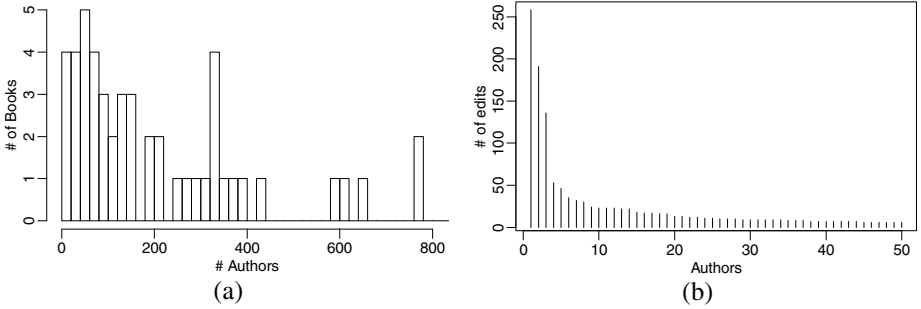


Fig. 2. a. Distribution of number of authors per book. Not included: 3 books with over 1000 authors (1111, 1039 and 2631). **b.** Distribution of number of edits by authors sorted in descending order (follows a zipf distribution).

Next, we consider the number n of contributors responsible for $x\%$ of the main edits. To compute this, we sort contributors from largest to smallest, based on the number of their edits, as in Fig. 2.b. Here, n is the smallest integer l for which the first l authors made at least $x\%$ of the edits. Note that for 57% of the books, one author is responsible for at least 25% of the main edits. The median number of editors responsible for 50% of the main edits is 3.5 and rises to 14 for 75% of the main edits. Interestingly, we did not observe any effect of the `RedundancyPatterns` on the total number of contributors, whether the number of contributors was responsible for 25%, 50% or 75% of the main edits.

For each book, we also computed the Gini coefficient [7], which indicates whether a book was written by a few lead contributors relative to their total number. As above, we did not observe any effect of `RedundancyPatterns` on this coefficient.

The Evolution of Wikibooks. Fig. 3 shows the distribution of the book’s age (in days) at the time it became featured (featured time). The age of featured books ranges from 114 to 2034 days with a 1st, 2nd (median) and 3rd quartile of 545, 904 and 1198 days, respectively, with a mean of 900.1 ± 446 days. We observed no effect of the `RedundancyPatterns` on this measure.

Wikibooks contain a varying number of separate pages, ranging from a minimum of 10 pages to a maximum of 646 pages (1st quartile = 26, median = 45, mean = 86 ± 113 , 3rd quartile = 90 pages). The number of words in a book ranges from 3945 to 525300 (1st quartile = 3278, median = 4981, mean = 71560 ± 79255 , 3rd quartile =

⁵ Zipf’s law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc.

80930 words). We did not observe any effect of the `RedundancyPatterns` on the number of words nor on the number of pages of a book.

When we examined the books, we observed only a marginal correlation between the number contributors (25%, 50%, 75%, and 100% of contributors) and (1) the age, (2) the number of pages, and (3) the number of words. The only significant correlation is between the number of contributors and the age of a book at featured time ($r^2 = 0.366$). More specifically, the age of a book is about $1.2 \times \text{Num of Contributors} + 632$ ($p < 0.0001$), after removing the three books with the largest number of authors. This suggests that over time more contributors became involved.

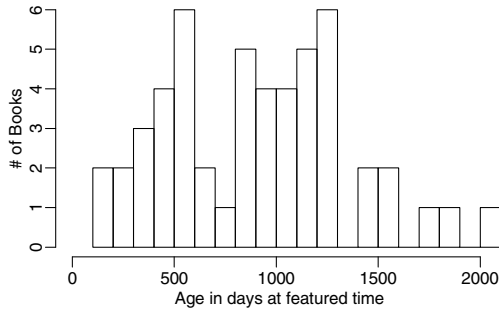


Fig. 3. Distribution of the age of books (in days) when books became featured

Surprisingly, we found no correlation between the age of a book (at featured time) and its number of words. We did, however, observe a correlation (slightly increasing) between the age of a book and the number of pages ($p = 0.0167$ with a low $r^2 = 0.093$), and a clear increasing correlation between the number pages and the number of words ($r^2 = 0.5945$, $p < 0.0001$).

5 Analysis 3: Collaborative Interactions

In traditional collaborations, careful planning is usually essential. However, in on-line collaborations, contributors are under no obligation to follow a plan or discuss issues with each other. Successful collaborative environments such as Wikipedia include both implicit and explicit coordination mechanisms to support collaboration [14]. Wikibooks offer two basic coordination mechanisms: *Predicted pages* specify the structure of the book and identify where additional writing is required. They act as an implicit coordination mechanism in which authors see broken links that point to yet-to-be-written pages and can contribute text accordingly. *Talk pages* enable authors to discuss content and negotiate changes, and act as an explicit coordination mechanism.

Research Questions: We are particularly interested in:

- What is the impact of communication and coordination mechanisms on the appearance of redundancy within the text?
- How does the authors' use of *predicted pages* (broken links) and *talk pages* affect redundancy patterns?

5.1 Measures of Explicit and Implicit Coordination

Every Wikibooks page has a talk page in which authors attach comments and discuss content. This enables them to organize the writing task and resolve disputes that may arise and also provided us with a quantitative measure of explicit coordination.

We also created a quantitative measure of implicit coordination based on the structure and number of predicted pages. We also counted the number of edits made by different authors. To differentiate between prolific and casual authors, we sorted authors according to their contributions. We then calculated the inter-quartile range and the Gini coefficient [7] of authors' contributions to measure the skew of contributions within each book. We applied these measures to talk pages and content pages separately.

5.2 Results

Implicit Coordination using Predicted Pages. We found a moderate correlation between the number of predicted pages and the number of pages in a book ($r^2 = 0.574$). The number of predicted pages represents about 19% ($p < 0.0001$) of the total number of pages. Based on this result, we consider the percentage of predicted pages as a finer measure of implicit coordination.

The `RedundancyPatterns` had a significant effect on the percentage of predicted pages ($p = 0.0101$). Of the total number of books, 14 had no predicted pages, eight belonged to *Initially-Redundant*, three belonged to *Suddenly-Redundant*, and three belonged to *Non-Redundant*. This effect can be observed in the box plot in Fig. 4. A Tukey test shows that the percentage of predicted pages is significantly greater for *Non-Redundant* books.

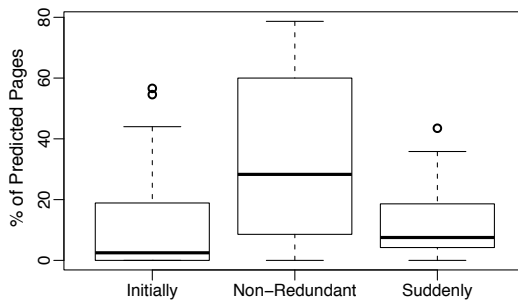


Fig. 4. Box plot showing the percentage of predicted page for each redundancy pattern

Note that we found no correlation between the percentage of predicted pages and the number of authors (contributions of 25%, 50%, 75%, and 100% of edits) and the number of words in a book. Finally, we only found a marginal decreasing correlation of this measure with the book age ($r^2 = 0.06404$).

Explicit Coordination using Talk Pages. Not all authors participated in talk page discussions. However, we did find a strong correlation between the total number of authors and how many participated in talk pages ($r^2 = 0.788$ for books with fewer than

1000 authors (see Fig. 5.a), and $r^2 = 0.923$ for all books). More specifically, about 13% ($p < 0.0001$) of authors participated in talk page discussions.

We found no significant effect of `RedundancyPatterns` on the number of authors into talk pages ($p = 0.079$). However, we found a significant effect of `RedundancyPatterns` on the number of authors for the 25% main talk edits ($p = 0.0325$). In both cases, *Suddenly-Redundant* books have more authors than the two other groups.

We examined the number of words in talk pages as a measure of participation in discussions. We observed that `RedundancyPatterns` has a significant effect on this measure ($p = 0.0171$). *Suddenly-Redundant* books have significantly more words in their talk pages than the other two groups (Fig. 5.b).

We also found a strong correlation between the number of words in talk pages and the time for the book to reach featured status ($p = 0.00123$, $r^2 = 0.177$). Note that we get a similar result (more discussion in *Suddenly-Redundant* books) even when we calculate the number of words in talk pages by day and when we normalize it by the number of pages and words in the book.

Finally, we examined the number of talk page edits. Again, *Suddenly-Redundant* books have significantly more talk page edits and a higher percentage of talk edits relative to main edits. Note that we find no correlation between the percentage of predicted pages and the number of words in talk pages (adjusted $r^2 = -0.003379$, $p = 0.366$). This suggests that these two measures are orthogonal.

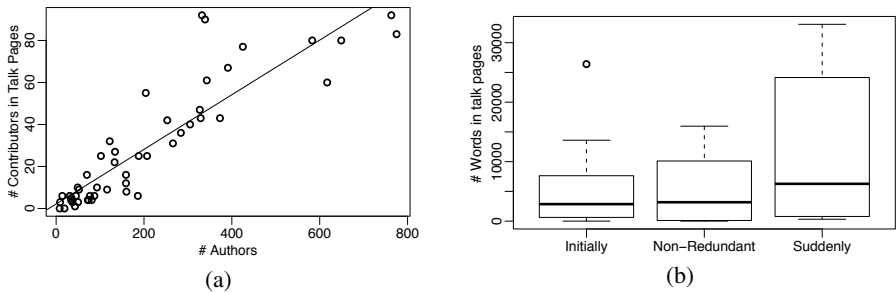


Fig. 5. a. Correlation between total number of authors and number of authors who participated in talk pages. **b.** Box plot of the talk page word count for each `RedundancyPatterns`.

6 Summary of Findings

In summary, we found that:

- Redundancy can be categorized according to three distinct patterns:
 1. *Non-redundant* books are well planned from the beginning and require fewer talk pages to reach high-quality status.
 2. *Initially redundant* books begin with high redundancy, which drops as soon as authors use coordination tools to restructure the content.

3. *Suddenly redundant* books display sudden bursts of redundancy that must be resolved, requiring significantly more discussion to reach high-quality status.
 - The majority of each book's content (50-75%) is typically written by a small number of lead authors, supported by a larger number of additional authors.
 - The *predicted pages* feature acts as an implicit coordination mechanism that does not reduce redundancy in the early phases of a book's development but appears to lower the overall effort of writing a book.
 - *Suddenly redundant* books require significantly greater discussion among authors, with a correspondingly high use of talk pages compared to the other two groups, in order to reduce redundancy.
 - The number of authors is not correlated with levels or patterns of redundancy, provided that proper coordination mechanisms are used.

7 Revisiting the Results through Examples

The above analysis showed general trends. This section examines several books in greater detail, including several that are outliers with respect to our statistical analysis.

7.1 Using Implicit Coordination to Avoid Redundancy

We have seen that books with no redundancy tend to have a high percentage of predicted pages. For example, the `CONTROL SYSTEMS` book has 65 predicted pages, which addresses the majority of the 75 pages in its featured form. The amount of redundancy for this book remains low throughout its history, even though it includes very little discussion (only 1905 out of 74,567 words).

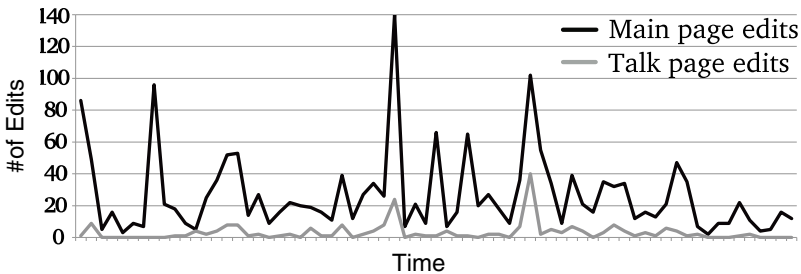


Fig. 6. `LATIN` book main pages (black) and talk pages (grey) edit history

The book includes a total of 93 authors with 10 contributors to the talk page discussion. Our analysis suggests that, although these authors engaged in relatively little discussion, they still managed to collaborate without producing redundant text.

Of the books that include no redundancy, not all make extensive use of predicted pages (see Fig. 4). These books achieve coordination through the use of talk page discussions. For example, even though the `LATIN` book included only two predicted pages of the 87 in its featured form, discussion of the book's structure via talk pages

occurred early in its development and continued throughout. Fig. 6 shows the history of main edits and talk page edits and the a strong correlation between them.

The talk page collaboration mechanism allowed 21 lead authors to coordinate the activities of 401 supporting authors on a variety of different topics. This is particularly evident in the talk page in which 77 authors actively discussed the status of the book.

7.2 Using Explicit Coordination to Remove Redundancy

Suddenly redundant books, with their sudden bursts of redundant material, rarely used on-line coordination mechanisms, either via predicted pages or talk pages. Similarly, in *Initially redundant* books, the redundancy curves rise sharply at the beginning due to lack of coordination. The most extreme example is *XFORMS*, with only one predicted page out of 154. For the first six months, *XFORMS* was developed mainly by one lead author (Gini coefficient $G=0.936$), as was *SPECIAL RELATIVITY* ($G=0.874$) (3 predicted pages out of 29).

Initial development of *ARIMAA* involved three lead and twelve supporting authors ($G=0.687$, 1 predicted page out of 53). In each case, the increase in redundancy is not related to the number of authors who contributed to the books, but rather is due to the authors' approaches to writing, in which they avoided planning the different sections of the book on-line. The three lead authors wrote separately without any interaction among themselves or coordination of future activities. As a result, the other twelve authors who contributed to the book made rather chaotic edits, adding pages that contained redundant information (see also Fig. 1.b).

Lead authors of 17 books compensated for the initial lack of planning by later performing a radical restructuring of the book, with a corresponding reduction in redundancy. For example, the lead author of *XFORMS* restructured the book into sections, whereas lead authors for both *SPECIAL RELATIVITY* and *ARIMAA* used a talk page to direct changes and amendments.

Another example is *C# PROGRAMMING*, a book in which redundancy increased when four new active authors began to contribute to the book independently, apparently paying little attention to previous contributions by other authors. At this point, the structure was not clear, authors' contributions were not well organized and included no discussion, with only 4 predicted pages and 13 pages of content. Two authors wrote extensively about similar concepts, unaware of each other's contributions.

Finally, one author noticed the overlap and used talk pages to discuss how to restructure the content. This major change in structure was accompanied by a subsequent plunge in redundancy (Fig. 1.c). Over the next months, 50 new authors began to contribute to the book, ultimately leading to an increase in the size of the book by 30%. This demonstrates the importance of structuring books in a logical way, so that authors have a clear idea of what each section should contain, without needing to read the full contents of the book.

7.3 Redundancy Might Be Beneficial

While redundancy within a book as a whole is undesirable, specific increases in redundancy can be beneficial if they lead to restructuring and result in a net decrease in

redundancy, e.g. the C# PROGRAMMING example. In some cases the detection of excess redundancy sparks a conversation about the book that generates interest and makes contributing content more appealing. We see this in all ten *Suddenly-Redundant* books and all eight *Initially-Redundant* books.

Another positive restructuring activity was seen in the EUROPEAN HISTORY book, in which a sudden spike in redundancy occurred when a new author contributed text in a new section, even though the same content was already present elsewhere. The other authors chose to retain the new text and re-integrate other relevant text into the new section.

However, sometimes adding text is simply a duplication of effort, as in FORMAL LOGIC in which a casual author increased the redundancy in the book by adding material to a page that was already covered in other pages. Redundancy decreased sharply after this text was removed by the main editor.

8 Discussion

Our analysis of collaboration within Wikibooks is consistent with previous research on Wikipedia with respect to the correlation between communication levels and quality. Like Kittur and Kraut [14], we found correlations between the use of implicit and explicit collaboration strategies, the distribution of authors and article quality. Overall, we found that despite the large number of contributors, most authorship, at least with high-quality ‘featured’ books, is concentrated among a few lead authors. On average, 75% of a book is written by no more than 14 authors, with a much larger group of supporting contributors who write the rest.

In addition, we observed the effect that collaboration activities have upon the appearance of redundant material within a book. When communication mechanisms are not properly used, authors tend to edit chaotically, which increases the quantity of redundant text. One might expect that the presence of large numbers of authors would lead to duplication of effort and highly redundant text. However, this need not be the case. The number of authors is not correlated with the presence of redundant text. Instead, duplication of effort occurs only when communication mechanisms are used improperly and contributions are chaotic.

Redundancy is normally viewed as a negative characteristic in a book, increasing the effort necessary for the book to become featured. We saw that books with highly redundant content required significantly more coordination effort, even though fewer authors contributed. However, the presence of redundancy can sometimes have a beneficial effect on the quality of the book. When lead authors become aware of redundant text, it triggers the restructuring of redundant pages and a discussion among the authors. In some cases, this can attract new contributors to the book, as seen in the books on C# PROGRAMMING, FORMAL LOGIC and US HISTORY. We do not have a direct measure of how aware authors are of the existence of redundancy, but suggest that lead authors would benefit from tools that draw attention to levels of redundancy since it will encourage them to resolve it, thus improving the quality of the book.

In general, our data suggests if lead authors set the direction, structure and scope of the book from the beginning, a variety of implicit and explicit coordination mechanisms can be used to support subsequent development, aiding future development.

As the book matures and coordination requirements ease, tasks may be more effectively distributed to a larger group of authors. This is consistent with other research that suggests that explicit communication through coordination is most beneficial in the early stages of the collaboration, when the structure is unconstrained [14].

Implications for Design: We found that planning of the book’s structure, whether implicit (as in predicted pages) or explicit (as in talk pages), has a positive impact upon future coordination of writing activities. Some books were carefully planned from the beginning and could be successfully managed with or without explicit coordination during subsequent writing phases. In contrast, books that were not initially well-planned suffered from a rapid increase in redundancy within the text, either at the beginning, or at different points during the book’s development. All of these books eventually ended up as high-quality books, with virtually no redundancy, but such books required significantly more communication and collaboration activities to compensate for the bursts of redundancy.

We believe that authors of large-scale collaborative writing projects will benefit from tools that support the communication and collaboration mechanisms by providing authors with visualizations of the following:

1. *Sections containing redundant text.* This would facilitate restructuring and merging changes, as well as potentially triggering participation by new authors.
2. *Number and types of edits to each page.* This would clarify the structure of each page and let authors focus on writing new sections, rather than extracting redundancy from previously written pages.
3. *Sections containing topics already covered.* This would help authors focus on writing relevant parts of the book instead of duplicating existing content.

9 Summary and Future Work

This paper examines how large groups of volunteer authors coordinated their activities in order to create and edit 51 ‘featured’ or high-quality Wikibooks. We used a combination of information retrieval and statistical methods to develop quantitative measures of coordination activities and identified several factors that are significantly correlated with effective collaboration.

One contribution is the use of redundant text as a measure of coordination effectiveness. Not only did we find that redundancy was inversely correlated with quality, but we also were able to identify different patterns of redundancy over time. *Non-redundant* books are highly coordinated from the very beginning, and progress smoothly to completion. *Initially redundant* books begin with little structure, but once the key authors identify the presence of redundancy and begin to take measures to reduce it, these books also progress smoothly towards completion. *Suddenly redundant* books are poorly planned in the beginning and suffer from spikes in redundancy over time, as people duplicate each other’s efforts. These books require by far the most discussion and late-stage coordination in order to become high-quality books.

A second contribution relates to the deeper understanding of the roles of authors in large-scale collaborative writing projects. We found that, even though it appears as

though thousands of authors have contributed, in practice, over 75% of the writing is produced by a small core group of lead authors. This implies that these authors need specialized tools for visualizing redundancy, which should help them to more effectively allocate writing tasks and better coordinate everyone's activities.

Our next step will be to modify our algorithms to enable authors to obtain successive snapshots of each level of redundancy. We plan to examine whether providing these authors with ways to visualize this information during the writing process can help to reduce development time and increase the likelihood of writing a high-quality book.

Acknowledgments. The authors of this research would like to thank Prof. Hal Abelson for inspiration and invaluable help in setting up this research idea and Theophanis Tsandilas for comments and suggestions on earlier drafts. Ching-Man Au Yeung worked on this research while at the University of Southampton.

References

1. Brooks, F.P.: *The Mythical Man-Month: Essays on Software Engineering*. Addison-Wesley, Reading (1995)
2. Chesney, T.: An empirical examination of Wikipedia's credibility. *First Monday* 11(11) (2006)
3. Chevalier, F., Dragicevic, P., Bezerianos, A., Fekete, J.D.: Using text animated transitions to support navigation in document histories. In: *Proc. CHI*, pp. 683–692. ACM, New York (2010)
4. Clearwater, S.H., Huberman, B.A., Hogg, T.: Cooperative solution of constraint satisfaction problems. *Science* 254, 1181–1183 (1991)
5. Emigh, W., Herring, S.C.: Collaborative authoring on the web: A genre analysis of online encyclopedias. In: *Proc. HICSS* (2005)
6. Encyclopedia Britannica Inc.: Fatally flawed: refuting the recent study on encyclopedic accuracy by the journal *Nature* (March 2006)
7. Gastwirth, J.L.: The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics* 54(3), 306–316 (1972)
8. Giles, J.: Internet encyclopedia as go head to head. *Nature* 438, 900–901 (2005)
9. Goldstein, J., Mittal, V., Carbonell, J., Kantrowitz, M.: Multi-document summarization by sentence extraction. In: *Proc. NAACL-ANLP*, pp. 40–48. ACL (2000)
10. Gutwin, C., Benford, S., Dyck, J., Fraser, M., Vaghi, I., Greenhalgh, C.: Revealing delay in collaborative environments. In: *Proc. CHI*, pp. 503–510. ACM, New York (2004)
11. Hill, G.W.: Group versus individual performance: are n+1 heads better than one? *Psychological Bulletin* 91, 517–539 (1982)
12. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data* 2(2), 1–25 (2008)
13. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: *Proc. CHI*, pp. 453–456. ACM, New York (2008)
14. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In: *Proc. CSCW*, pp. 37–46. ACM, New York (2008)
15. Kittur, A., Lee, B., Kraut, R.E.: Coordination in collective intelligence: the role of team structure and task interdependence. In: *Proc. CHI*, pp. 1495–1504. ACM, New York (2009)
16. Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He says, she says: conflict and coordination in Wikipedia. In: *Proc. CHI*, pp. 453–462. ACM, New York (2007)

17. Kowalski, G.: *Information retrieval systems: theory and implementation*. Kluwer Academic, Dordrecht (1997)
18. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. *Discourse Processes* 25, 259–284 (1998)
19. Lerner, J., Pathak, P.A., Tirole, J.: The dynamics of open-source contributors. *American Economic Review* 96(2), 114–118 (2006)
20. Li, L., Zhou, K., Xue, G.R., Zha, H., Yu, Y.: Enhancing diversity, coverage and balance for summarization through structure learning. In: *Proc. WWW*, pp. 71–80. ACM, New York (2009)
21. Li, Y., McLean, D., Bandar, Z.A., O’Shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowl. Data Eng.* 18(8), 1138–1150 (2006)
22. Lih, A.: Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In: *Proc. ISOJ*, pp. 16–17 (2004)
23. Mackay, W.E.: Patterns of sharing customizable software. In: *Proc. CSCW*, pp. 209–221. ACM, New York (1990)
24. Nardi, B.A., Miller, J.R.: Twinkling lights and nested loops: distributed problem solving and spreadsheet development. *Int. J. Man-Mach. Stud.* 34(2), 161–184 (1991)
25. Panciera, K., Halfaker, A., Terveen, L.: Wikipedians are born, not made: a study of power editors on wikipedia. In: *Proc. GROUP*, pp. 51–60. ACM, New York (2009)
26. Pedersen, T., Patwardhan, S.: Wordnet:similarity - measuring the relatedness of concepts. In: *Proc. AAAI*, pp. 1024–1025 (2004)
27. Raymond, E.S.: *The Cathedral and the Bazaar*. O’Reilly, Sebastopol (2001)
28. Sajjapanroj, S., Bonk, C.J., Lee, M.M., Lin, M.F.: The challenges and successes of wiki-bookian experts and Wikibook novices: Classroom and community collaborative experiences. In: *Proc. AERA* (2007)
29. Steiner, I.D.: *Group process and productivity*. Academic Press, London (1972)
30. Stewart, G.L.: A meta-analytic review of relationships between team design features and team performance. *Journal of Management* 32, 26–55 (2006)
31. Thagard, P.: Collaborative knowledge. *Nous* 31, 242–261 (1997)
32. Turney, P.D.: Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) *ECML 2001*. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
33. Viégas, F.B., Wattenberg, M., Dave, K.: Studying cooperation and conflict between authors with history flow visualizations. In: *Proc. CHI*, pp. 575–582. ACM, New York (2004)
34. Xiao, Y., Baker, P.B., O’Shea, P.M., Allen, D.W.: Wikibook as college textbook: a case study of college students’ participation in writing, editing and using a wikibook as primary course textbook. In: *Proc. AERA* (2007)
35. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: *Proc. Machine Learning*, pp. 412–420 (1997)
36. Zipf, G.K.: *The Psychobiology of Language*. Houghton-Mifflin, Boston (1935)

Towards Interoperability in Municipal Government: A Study of Information Sharing Practices

Stacy F. Hobson, Rangachari Anand, Jeaha Yang, and Juhnyoung Lee

IBM T. J. Watson Research Center,
Hawthorne, New York, USA
{stacypre, ranand, jeaha, jy1}@us.ibm.com

Abstract. Municipal governments rely heavily on the sharing of data between departments as a means to provide high-quality and timely service to its citizens. Common tasks such as parcel renovations require the involvement of multiple departments such as Building, Planning, Zoning, Assessment and Tax to achieve the ultimate goals. However, the software applications used to support the work of these departments are provided by independent software vendors and are not integrated with one another. Therefore, municipal employees rely heavily on manual methods for data sharing. We conducted a study of 12 municipal governments to understand their information sharing needs and practices. We focused on the interaction and information sharing within and between municipal departments. Our findings can be used to shape future research on e-government initiatives and interoperability of municipal applications.

Keywords: Information Sharing, Cooperative Work, Municipal Government, e-government.

1 Introduction

Citizens rely on municipal governments for the provisioning of services that cannot be achieved by an individual alone. These services include protection, maintenance of public areas and highways, environmental planning, safety, and governance. To this end, municipalities organize their internal department structure around specific service sub-areas. Departments such as police, fire, emergency services, and justice are responsible for services under the domain of protection. The sharing of information is essential for workloads that are divided across multiple people, departments and disparate applications. Information sharing between departments is used to help manage resources, communicate essential information about citizens and services, support the provisioning of services, and is useful in crime prevention efforts.

Departmental employees are expected to collaborate with other departments as needed, e.g. as mandated by policy or laws or in support of citizen-based services that span multiple departments. However, the information technology structure prevalent in most municipal governments does not easily support information sharing practices. IT applications that are interoperable can help in reducing employee workload, and also the expenses associated with and errors stemming from the manual transfer of data.

Hoogenveen points out that municipal police departments often have database systems that are central to their work and used by their personnel, but lack an efficient manner to share information with other departments [12]. Atabakhsh supports this assertion and states that the necessary tools for data retrieval, filtering, integration and presentation have not been sufficiently developed [1].

Our research goal is to identify methods and design technologies that will increase employee efficiency and help reduce costs in the municipal government domain. As part of this, we noted how the intersection of IT applications, organizational structure, aspects of social interactions and internal policy all worked together to influence the level of collaboration and information sharing between municipal departments. To this end, we conducted a year-long study of 12 municipalities, focusing on the working practices of employees, the use of IT applications as work aids, the associated manual and IT-supported flow of information, the departmental structure and division of service responsibilities across municipal departments.

1.1 Related Work

Information sharing is a common topic of discussion in the HCI-related literature. There have been extensive studies on information sharing practices in the private sector, e.g. within enterprises and between independent groups [2, 10, 17, 20]. Research on information sharing in the public sector have largely been limited to criminal justice [1, 23] and health and social service organizations, e.g. medical services [25], public health agencies [1, 23], and homeless shelters [18]. These and other studies on information sharing in the public sector relate to interaction between *independent* agencies [7, 14, 16, 19]. The division of labor and need for collaboration in the public sector can occur at many levels, e.g. at the local departmental level, at a peer agency level (i.e. from one municipality of government organization to another), from an agency level to another agency lower or higher in the hierarchy (from municipality to county or state agency), and from one nation to another. We maintain that some of the issues prevalent in information sharing initiatives *between* government agencies are also present *within* municipalities, yet the nature of municipal government organizational and IT structures can contribute to additional barriers to information sharing. This discussion is detailed in Sections 2 and 3.

Other related works contribute to the discussion of the role of IT in public sector information sharing practices. In an extensive research report by Harvard University's Center for Business and Government [9], the authors state that despite advancement in IT the government domain, *stovepipes* continue to dominate. Stovepipes are defined as "an inability to communicate across boundaries, between bureaucratic organizations or databases, due to lack of interoperability across hardware, software, or data systems; professional and cultural norms that prohibit or discourage information sharing; or legal strictures against communication." The authors also report that IT can be used to facilitate information sharing between entities and holds extraordinary promise as a vehicle for combating stovepipes, but that additional study is necessary to illuminate other factors influencing information sharing, such as aspects related to social science, policy, and trust research.

Governments are also becoming increasingly interested in electronic government (or e-government) as a means to increase the quality of service to its citizens and

support transparency in government while maintaining or even reducing costs. As part of e-government initiatives, municipal websites can be developed and leveraged to allow citizens to access forms, submit service requests, and check the status of their request online, rather than in-person, thereby reducing the workload of municipal employees. Interoperability of IT systems is reported as a critical factor for (successful) e-government initiatives [7, 15, 24, 26], since the information that is needed for municipal websites stems from multiple internal departments and must be consolidated, parsed and presented to the citizen as needed. Kaylor et. al [13] point out that municipalities are very interested in implementing e-government but that they may have to pick and choose which subset of services to implement because of their limited funding to support e-government initiatives and the high costs of the related technologies. The necessity of automatic sharing of data and interoperability between systems is a contributing factor to the significant cost of e-government technologies. Caffrey suggests that governments can be even more efficient and effective just based on integration and use of the information they already collect and maintain [4].

In [3], researchers at the University of Albany's Center for Technology in Government suggest that a number of factors are critical for governments to get full value out of the information that they collect, create and maintain. Two of these factors are integration of systems and the proper availability and use of information. In [5, 6, 8] the researchers also describe some of the benefits governments will realize through the integration of systems and availability of information, which include help with program planning, service evaluation, decision making and delivery of services to citizens.

Since the service offerings, technology needs and IT innovations of the public sector differs greatly from that of the private sector [29], we chose to study information sharing in a critical yet often overlooked area of the public sector, municipal governments.

Norman states that computer systems used to help people must be built to fit the needs of the people, and that systems cannot work for collaborative groups unless there is a deep fundamental understanding of the people, groups and how they work [22]. In the municipal government domain, we noticed that many of the inherent disparities of the IT application offerings makes information sharing a unique challenge for municipal employees and do not adequately help the employees collaborate. However, since the provision and management of services involve multiple departments, collaboration and information sharing is critical to the employees' work. This motivates our interest in studying the current ways municipal employees collaborate, share information and use IT applications. It also supports our interest in designing technology solutions to better support work at the municipal level.

We extend the research on information sharing in the public sector by providing an examination of the practices *within* public sector organizations, specifically municipal governments, in the remainder of the paper and compare our findings with that of the prior research.

2 Study Methodology

We conducted a study of 12 local governments in New York State over a 14 month period. The governments included three cities, four towns, four villages and one county. We employed semi-structured group and individual interviews, discussions

with municipal administrators and multiple direct observation sessions of the employees as they carried out their daily routines. The demographic makeup of the municipalities varied greatly; ranging from urban areas with a median household income of \$29,000 to suburban areas with a median household income of \$160,000. The populations of the municipalities ranged from approximately 10,000 to 65,000 people. The demographics of the governments were found to be related to the IT inventory assessment; we noted that municipalities with higher income levels or high populations had more modern IT systems and software, while municipalities whose demographics both fell in the low- to mid-range utilized more paper-based manual processing for record keeping and service management.

Although our study was limited to a few medium-sized municipalities in New York State, it has been reported by US government census data that municipalities in this state, and the northeast area of the United States are similar in power, structure and function [28], we believe that these findings may be extendable to other municipalities in the same size range throughout the state of New York and other neighboring states.

During the study we aimed to determine how services were provided and managed by the municipality including the division of service responsibilities across departments, and how information technology (IT) systems and paper documents were used to support service management. We focused on the use of IT software by employees to carry out their work, inter-departmental information sharing practices and we engaged in a detailed analysis of paper documents, especially those generated and used in the transfer of information from one department to another.

2.1 Study Details

The study was qualitative in nature, and employed direct observations of the employees as they carried out their day to day work along with semi-structured individual and group interviews. Study participants were identified and organized through municipal associations and conferences based on their willingness to be studied. Seventy-one employees were involved in the study, representing fifteen of the seventeen municipal departments¹ (listed under the administrator) from Figure 1.

We began our studies by interviewing the administrator for the municipality. The interview questions focused on determining the administrator's view of the organization, the areas of concern with respect to collaboration and information sharing between departments, and his/her identification of the departments in which their work and the need to share information was deemed most critical. Some of the questions used in the interview included:

1. What are your short-term goals for your organization?
2. What are your long-term goals?
3. How is the measure of performance of the overall organization measured?
4. Are there different measures for department-level performance?
5. Describe some examples of information that is communicated between departments.

¹ Very few municipalities studied had an animal control department since this function was often handled by the police. Similarly, in most cases, the highway department function appeared under the department of public works. Both areas were determined to be independent departments in municipalities much larger than the ones we studied.

6. How is this information currently communicated?
7. Is there a specific department that has a primary responsibility for this information, or is the responsibility shared between departments? (*asked for each example*)
8. What is your primary concern with communication between departments?
9. What are your additional concerns?
10. Which points of communication between departments are most critical to *<the previously identified performance measure>?*
11. Which points of communication are most critical to the delivery of services to your citizens?

After the interview of the administrator, we did initial observations of the employees from various departments. We observed and interacted with as many departmental employees as the administrator, or the coordinator of our visit would allow. The employees were told that we were interested in understanding municipal operations and their views on the work aids available and used to help them carry out their work. We chose to introduce our study to the employees in this manner to try and reduce concerns prompted by our visit (e.g. concerns of possible downsizing or critical inspection of their work performance).

During our initial observations of their work, we asked general questions about their responsibilities, challenges in their work, work methods, use of technology, to understand the scope of their work. We also allowed the employees some leverage in steering the direction and duration of this part of the study, in an effort to engage their trust and not impede their normal work.

After the first observation, we interviewed the employee using questions similar to those asked of the administrator, with more focus at the department level and appended by questions related to technology and work aids. After this interview, we asked the employees if we could continue to watch them as they worked without interrupting them, to get a better understanding of the internal operations. During this observation, we focused on points of interaction or overlap with other employees related to information or knowledge requests. For example, we noted when employees called or received calls from other employees, when they emailed or printed information to give to other employees, and the type of information exchanged and/or requested. Separately, we also engaged in some of the social interactions between employees (we were fortunate to be invited to a few office parties and events) to determine whether work items were discussed and what work-related topics of interest were emerged in the conversations. The goal of the extended observation was to determine areas of information sharing that were not identified by the employees in the interviews, and to better understand the scope of the information sharing.

Detailed notes from the interviews and observations were taken by the study team, and copies of work items were obtained. These work items included the common paperwork and forms that were processed for each department, the reports that were generated, and the objects that requested by, transferred to and received from other departments. Confidentiality was not a significant concern since most information is covered under New York State's Freedom of Information Law (FOIL) [21]. Items that were deemed confidential or of a sensitive nature (e.g. personnel records, birth records, etc) were discussed but not viewed or obtained.

2.2 Organizational Structure

During our study, we examined the organizational structure of each the municipalities and found that a similar basic structure emerged. The municipalities, regardless of size, provided a base set of services to its citizens and generally utilized a common departmental structure to support these services. We note that although local, county or state law may mandate the offering of certain services or the existing of specific employee positions, departmental structure is not mandated by law. Figure 1 shows the example of the common municipal structure we discovered. The departments highlighted in bold were present in all the municipalities, and as such, serves as the basic departmental structure. This basic structure contained only the departments that are most critical to municipal operations and included other service areas as sub-functions supported within the primary departments. The assignment of the sub-functional service areas to specific departments varied more than the primary organization itself; the sub-functional areas depicted in our figure represent one of the examples we encountered.

The six departments included in the base structure most often were present in small municipalities that did not serve a large population (< 30,000). Larger municipalities (population > 30,000) tended to have a more expansive structure, which corresponds to all seventeen departments listed in the figure.

We must note that all municipalities do not follow this structure exactly; rather, this is the general organization that has emerged from the municipalities we studied and is used for service delivery to its citizens. Individual municipalities may use different names for the department, may combine two departments into a single department, or

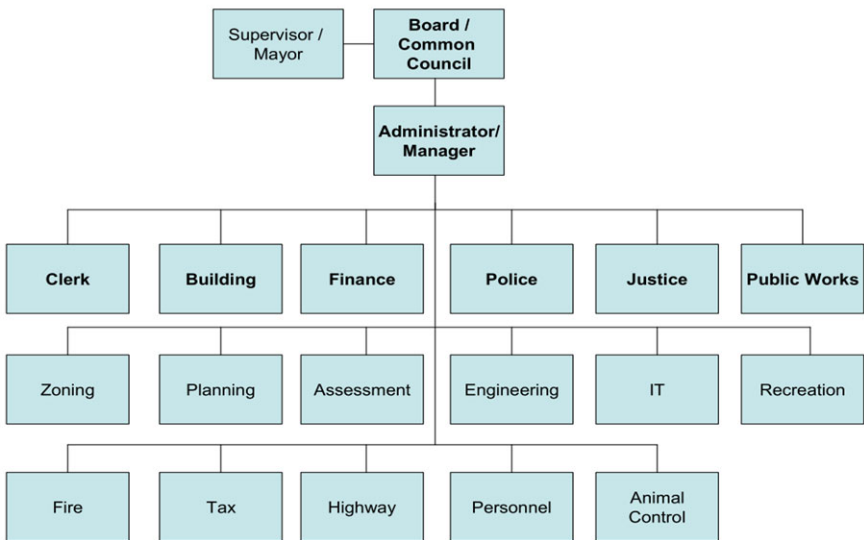


Fig. 1. Common Municipal Organization Structure

vice versa, but the department's operational responsibilities and associated work tasks primarily align with the figures illustrated here. For example, the services provided by the *Public Works* department were available in each municipality, although the specific department name in a single municipality may have been *Water and Sewer*. We also note that the organization may have small differences stemming from municipal type. For example, while villages and cities have Fire Departments under their organizational structure, towns do not. Towns may organize special districts (outside of the town's internal organizational structure) to provide fire fighting services.

2.3 Association of Organizational Structure and IT Applications

Municipalities offer a range of services to its citizens, including protection, maintenance of roads and public areas, sanitation, and support programs. The internal departments are responsible for sub-sets of the service area offering and management functions. Seven types of service offerings are listed below along with the departments who jointly provide the services:

- Protective services – Police, Fire, Justice, Animal Control
- Sanitation/utilities – Public Works, Zoning, Engineering
- Environmental services – Public Works, Zoning, Engineering
- Support programs – Recreation, Clerk
- Internal programs – IT, Finance, Personnel
- Parcel management – Engineering, Building, Planning, Assessment, Tax, Clerk
- Personal services – Clerk, Justice, Police

Most of the IT applications used to support service delivery align directly to the departmental structure, with a single or multiple application(s) covering only the tasks associated with one department. However, services offered and provided to citizens often involve the work of many departments. Consider, for example, a parcel that has been recently renovated by its owners. The parcel owner applies for a building permit from the Building department. After the work has been completed and the Certificate of Occupancy has been issued, the Assessor's office must update the parcel details to reflect the renovations and initiate a parcel re-assessment. The newly assessed value has to be updated in the tax system for calculation of the property tax liability. Finally, the property tax payments collected by the tax department are directly entered into the tax system, and must later be updated in and reconciled with the accounts of the Finance department's General Ledger. This example describes the complex interactions between four departments as part of the administration for parcel management services including the renovations, assessment and property tax processes.

In many cases, the departments all use distinct, non-integrated, custom software programs to aid in their work. Additionally, departments may also use paper-based records and manual record-keeping. Figure 2 below shows the specific departmental and IT application structure from one of the municipalities we studied. Note that there are 10 different IT systems for the 8 departments, and that one additional department achieves their tasks entirely without the use of custom IT software (only through use of basic spreadsheet software and paper-based documentation methods).

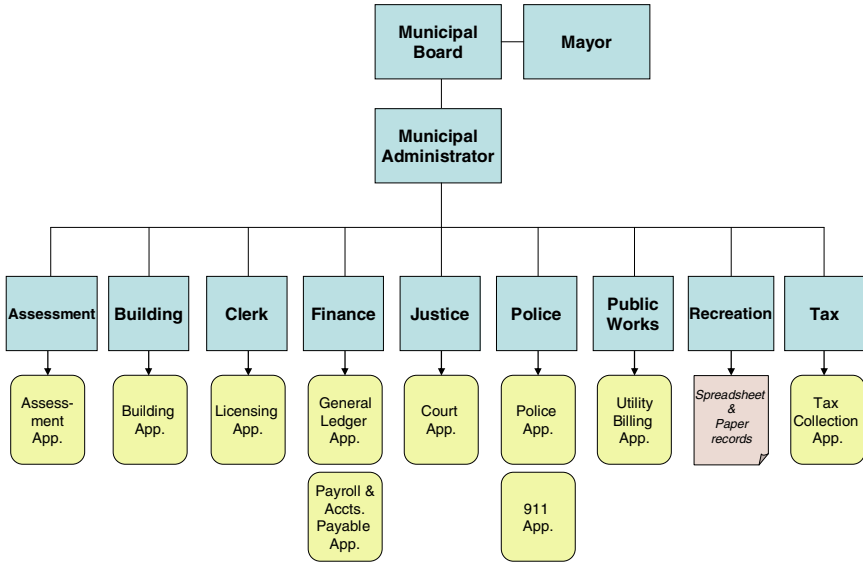


Fig. 2. Association of Department and IT applications

3 Information Sharing Practices

We introduced the notion of a common municipal structure and presented an example of the association of department and IT application in the previous sections as an introduction to the discussion on service boundaries. The types of information shared between departments can be static or dynamic. We define static information as information that is required and obtained by multiple groups, but remains unchanged in its representation or very minimally changed in its representation but unchanged in its function. We associate dynamic information with boundary objects. Boundary objects have been defined in the literature as objects that span multiple *social worlds*, are adaptable to satisfy the needs of each of these groups, while maintaining a common identity [27]. We consider departments within a municipality basic instances of social worlds. Examples of both types of information are described below.

Static information includes information at the data attribute level. For example, parcel contact details such as the mailing address, contact phone number, and owner name are items that often originate in a department such as building, assessment, or tax and must be shared throughout a large number of departments (building, zoning, assessment, tax, clerk, recreation, police and public works). Similarly, details related to property tax payments specifically amount due and/or payment amount, are communicated between the tax and finance department for accounting purposes, the tax and public works department for billing purposes (as a single bill for normal billing, or combined bill when a payment is overdue), and tax and recreation department for determination of eligibility requirements for certain recreation programs.

In this case, boundary objects would include the information that originates in one department, and is passed on to other departments for additional viewing, processing, appending, and archiving. Specific examples include a building permit or renovation request, which is created by a user and internally originates from the building or zoning departments. This document is then shared with departments such building or zoning (whichever is not the originating department) and planning for purposes related to the evaluation and determination of the appropriateness of the renovation plans, the assessment department as the initializer of what then becomes the assessment, and then the tax department for the detailed tax liability and billing document.

Another example is seen with the communication documents that flow between the fire, police, emergency services (although external) and/or the justice departments. The boundary object in this case would be the incident or case report, the object that is used to document an event that has associations with public safety, a possible law-breaking act, and/or in some cases, medical situations. Accuracy and timeliness in the transmission of this information is often critical for investigations, prosecution, and the prevention of future incidents or acts.

There are numerous challenges in the current methods for sharing of both types of information. An IT-specific challenge is related to the silo-ed nature of municipal applications (as described in the previous section). Significant manual work is often required to enable the transmission of information between departments when the IT does not support automatic sharing, and manual re-processing and re-keying of data can lead to omissions or data entry mistakes. Mistakes can have varying consequences from the extra time spent for identification and correction, up to a public safety disaster or threat. A deeper dive into some of the issues relating to information sharing within a single municipality is described below.

3.1 A More Detailed Information Sharing Example

Departments such as Police and Justice, Assessment and Tax, Personnel and Finance rely heavily on communication of data that is central to their work. The Police department must notify the Justice department of arrest and investigation details as input to current and future court cases. The Assessment department must periodically give the Tax department the *tax roll*, a listing of all taxable properties, their current assessment values, and valid exemptions. The Personnel department (and/or individual departments) maintains information on employee pay rates, raises, benefits, and work hours and these factors are used to calculate semi-monthly pay to produce payroll checks.

In one municipality, we found that 18 distinct pieces of information originating in the Tax department were shared with the Finance department for documentation and re-processing. Some of the information included printed records from the Tax application, documents that were generated in Microsoft Word or Excel as calculations data, and copies of bank slips for the tax payment deposits. Figure 3 shows the specific stack of information that the Tax department gives to the Finance department on a monthly basis. Much of this information is re-entered into the Finance application and then used for the process of monthly reconciliation with the bank statements. Figure 4 is a picture of the desk of the Finance department's manager, who is responsible for the reconciliation process. The three employees responsible for information sharing between the Tax and Finance departments spent about four total days per

Fig. 3. Printed Records Transmitted from Tax to Finance Department

month (one day for each of the two tax employees, and two days for the finance employee) directly involved in this process; producing, transmitting, re-keying and correcting the associated data.

3.2 Additional Factors Influencing Information Sharing

So far, we have focused on the importance of intra-departmental information sharing in municipalities, and briefly described how the non-interoperability of IT applications influences the employees' methods of information sharing. We must also point out that municipalities may choose to pay for integration between two software programs or purchase pre-integrated software packages, especially for departments who share information frequently. This integration is useful in that it may reduce the amount of manual information sharing and processing that occurs and can allow for information to flow from one department to another automatically. However, the integration normally occurs between a set of applications and incurred very high costs for each pair (ranging from \$8,000 to \$20,000). This cost made it difficult for most of the municipalities we studied to afford more than a single integration.

We also determined that the organization of IT software can be a limiting factor for service offerings. During our study, a municipal administrator commented how information from the Clerk's department, e.g. animals registered to a specific parcel, can be useful to members of the Police and Fire departments when they are dispatched to the parcel for an investigation or emergency services. Currently, the primary way to have this information reflected in the Police or Fire department



Fig. 4. Desk of Finance Employee with Records Awaiting Processing

software is through manual transmission and data entry. Since municipal workers are often overburdened with their general responsibilities, it is unlikely for this to occur on a regular basis, if ever.

In addition to the availability and usage of IT applications as an influence of information sharing practices, there are a number of other factors that may have an impact. We encountered a few cases where applications that included pre-integrated modules for multiple departments were evaluated and were either not chosen at all or only purchased for a subset of the departments. We discovered that this occurred for a number of reasons:

A preference for different non-integrated applications for each department. The department employees liked to have autonomy in choosing the applications that they used. An application was often chosen based on its suitability, familiarity, or ease-of-use for a single department, with less emphasis put on their integrative capabilities. In one municipality, a finance department director and tax department director both recommended different pre-integrated software packages that covered both departments. However, the ones recommended were most suitable to the recommender's department and not as well tailored for the other department. The directors ultimately each chose the separate module of the system that was self-recommended and continued to manually transfer their information.

A resistance to change or preference for current manual methods for sharing. An employee or multiple employees tried to prevent the purchase of packages that would allow electronic means for information sharing. We interviewed a director of the

Justice department in one municipality and she stated that she “specifically chose a package that would not integrate with the police department’s application because [she] wanted them to continue to provide information to and request information from [her] department as they were used to.” (Additional analysis of the interviews from employees of both departments suggested an ongoing power struggle between the two departments.)

Cost as a prohibitive measure. He cost of the integrated application packages provided by a single vendor was higher than that of the individual applications provided by multiple differing vendors, or a preference to only purchase one application module from an integrated package and continue using existing applications for other departments.

We also note that additional social and personal factors such as trust, willingness to share knowledge, political beliefs and/or the existence of policies or laws on information sharing may all have varying limits of involvement.

4 Discussion

Many of the related literature described in Section 1.1 focus on information sharing practices *between* government organizations (inter-organizational information sharing) or information sharing in a single service area (like public health, public safety and social services). Our study has focused on information sharing *within* public sectors organizations, specifically, in the municipal government domain. In the previous sections, we described IT application silos as one of the barriers to easy and effective information sharing within municipal governments. We maintain that the cause of and suggested solutions to application silo issues in municipal government differ than that of other areas. It is understandable that a challenge with inter-organizational information sharing is caused by silo-ed IT applications because the participating organizations are distinct and independently run, and are more likely to purchase IT applications without consulting one another. However, municipal governments are primarily governed by a single body (a mayor, supervisor, or board/council) and have motivations for intra-organizational information sharing (improvements in quality of citizen-based services and policies/dictates by management). Yet, municipal departments are still able to function very independently and make decisions that further inhibit information sharing efforts.

Additionally, domains like public health, public safety and social services are usually discussed with an inter-organizational information sharing focus. The discussions that do focus on intra-organizational cases occur at levels higher than municipal governments. For example, in the US, social services and public health offerings are county or state-based and not handled by municipal governments (with the exception of very few extremely large cities like New York City). Significant challenges for public health, social services and other higher level government organizations include information privacy and strict policies on how and with whom data can be shared, and also where data is stored. This is less of a concern in municipal governments in that a great deal of the information (property tax details, assessments, building records, etc.) fall under the scope of Freedom of Information Laws [21] and therefore are publically accessible.

4.1 Implications for Future Systems

There are many challenges to interoperability in government. Two of the biggest barriers are time and funding. Municipal employees are often overburdened with their normal day-to-day duties and lack the necessary time needed to do an introspection of their information sharing needs and capabilities. When information sharing through technology is desired, municipalities often lack the extensive funding needed to enable point to point integration of their existing systems. Additionally, in the limited cases where software vendors offer “pre-integrated” technologies or software that span multiple departments, the span usually does not encompass all the departments involved in service management to the citizens. A distinct example we noted in our study was related to a new software package for the use in the Building, Planning and Zoning departments. The software provided a single point-of-use for all three departments and limited the need for manual information sharing between the departments, two significant issues were noted. The first was that the employees had difficulty in becoming proficient in using the system. The system was so vast and many menus and options that were available were not relevant to the work of a single department. Secondly, as noted above, since the citizen-based services related to parcel renovations spanned six separate departments (Building, Planning, Zoning, Assessment, Tax and Finance), manual information sharing was still needed to provide details to the remaining three departments (Assessment, Tax and Finance). We do note that the integrated system did reduce the need for manual processing within the three departments it covered, but did not wholly solve the information sharing needs related to the parcel renovations process.

Another distinct challenge lies in the view of providing a system that integrates all the information residing in the municipality and makes it available to the user. This solution is not optimal in that it overwhelms the user with too much information, only some of which is applicable to their work. Although this would help reduce the need for manual information sharing, it increases the cognitive load of the user and the amount of time spent parsing the information to find the details that are relevant to them. Also, information sharing issues can occur within a single department (e.g. data access or data migration issues) because of the great reliance on outdated IT systems and manual work methods in municipal governments. Therefore, IT projects or initiatives that aim to address information sharing problems through comprehensive data integration must also consider the need to gain access to records that are stored in the old systems or in paper format, which presents its own challenges.

A final challenge is in enabling the solution to fit the needs of each government. Since the set of IT applications in each government differ greatly, any solution that is offered to enable automatic integrations must be flexible so that it does not require significant individual customizations, yet remain economically feasible.

5 Conclusion and Future Work

Based on this study, we are currently designing a technology that will use a cloud-computing platform and hub-based integration capabilities to deliver municipal government IT services. We aim to improve municipal operations through increased quality of service and work efficiency from automatic information sharing capabilities, cost reductions from reduced labor, and the creation of advance analytic capabilities

available through integrated data. The hub-based integration capability differs from the traditional pair-wise integration capabilities currently possible within municipal governments. While pair-wise integration attempts to generate a component or add-on to link specific pairs of applications, our idea leverages the use of an information hub that coordinates data and event sharing among all the applications. When a new application is added to the municipal application set, it only needs to follow the APIs and guidelines that the information hub provides, and it can then be integrated with all the other applications (as allowed through configurations by a municipal system administrator). The pair-wise integration technique is often ad-hoc and costly; our hub-based integration design is aimed to be automatic, easily configurable and much less expensive. The cost can be down from N^2 (the expected cost for pair-wise integration across all applications) to N . More detail on this approach is available in [11].

This paper has helped to highlight the importance of intra-departmental information sharing in municipalities. Specifically, it is critical to the timely achievement of work, quality delivery of citizen-based services, and reduction of labor and cost associated with data re-entry. Today's significant reliance on manual methods for information sharing can often lead to mistakes that negatively impact the organization. Our illumination and discussion of these issues can be used as a foundation for researchers and key stakeholders of government initiatives to design alternative methods for information sharing for municipal organizations.

References

1. Atabakhsh, H., Larson, C., Peterson, T., Violette, C., Chen, H.: Information Sharing and Collaboration Policies within Government Agencies. In: Chen, H., Moore, R., Zeng, D.D., Leavitt, J. (eds.) *ISI 2004. LNCS*, vol. 3073, pp. 467–475. Springer, Heidelberg (2004)
2. Barua, A., Ravindran, S., Whinston, A.: Effective Intra-organizational Information Exchange. *Journal of Information Science* 23(3), 239–248 (1997)
3. Bloniarz, P., Canestraro, D., Cook, M., Cresswell, T., Dawes, S., LaVigne, M., Pardo, T., Scholl, J., Simon, S.: *Insider's Guide to Using Information in Government*. Center for Technology in Government, State University of New York at Albany, <http://www.ctg.albany.edu/static/usinginfo/index.htm>
4. Caffrey, L.: *Information Sharing Between and Within Governments*. Commonwealth Secretariat, London (1998)
5. Center for Technology in Government, State University of New York at Albany. *Dealing With Data Seminar Summary*, http://www.ctg.albany.edu/publications/reports/what_rules_govern/what_rules_govern.pdf
6. Center for Technology in Government, State University of New York at Albany. *What Rules Govern the Use of Information?*, http://www.ctg.albany.edu/publications/reports/dealing_with_data/dealing_with_data.pdf
7. Dawes, S.: Interagency information sharing: Expected benefits, manageable risks. *Journal of Policy Analysis and Management* 15(3), 377–394 (1996)
8. Dawes, S., Cresswell, A., Pardo, T.: From “Need to Know” to “Need to Share”: Tangled Problems, Information Boundaries, and the Building of Public Sector Knowledge Networks. *Public Administration Review* 69(3), 392–402 (2009)
9. Fountain, J.: *Information, Institutions and Governance: Advancing a Basic Social Science Research Program for Digital Government*. Harvard University Press, Cambridge (2003)

10. Goh, S.: Managing Effective Knowledge Transfer: an Integrative Framework and Some Practice Implications. *Journal of Knowledge Management* 6(1), 23–30 (2002)
11. Hobson, S., Anand, R., Yang, J., Lee, J.: Municipal Shared Services Cloud. To appear in Proceedings of the 2011 SRII Global Conference. IEEE Press, Los Alamitos (March 2011)
12. Hoogenveen, M., van der Meer, K.: Integration of Information Retrieval and Database Management in Support of Multi-media Police Work. *Journal of Information Science* 20(2), 79–87 (1994)
13. Kaylor, C., Deshazo, R., Van Eck, D.: Gauging e-government: A report on implementing services among American cities. *Government Information Quarterly* 18, 293–307 (2001)
14. Koch, M., Steckler, N., Delcambre, L., Tolle, T.: Examining Information Sharing Across Federal Agency Boundaries. Presented at the National Workshop on Digital Government, Kennedy School of Government. Harvard University Press, Cambridge (2002)
15. Klischewski, R., Scholl, H.: Information Quality as a Common Ground for Key Players in e-Government Integration and Interoperability. In: Proc. 39th HICSS. IEEE Press, Los Alamitos (2006)
16. Layne, K., Lee, J.: Developing fully functional e-government: A four-stage model. *Government Information Quarterly* 18, 122–136 (2001)
17. Lee, H., So, K., Tang, C.: The value of information sharing in a two-level supply chain. *Management Science* 46(5), 626–643 (2000)
18. Le Dantec, C., Edwards, W.: The View from the Trenches: Organization, Power and Technology at Two Nonprofit Homeless Outreach Centers. In: Proc. CSCW 2008, pp. 589–598. ACM Press, New York (2008)
19. Le Dantec, C., Edwards, W.: Across boundaries of influence and accountability: The multiple scales of public sector information systems. In: Proc. CHI 2010, pp. 627–636. ACM Press, New York (2010)
20. Li, J., Sikora, R., Shaw, M., Tan, G.: A strategic analysis of inter organizational information sharing. *Decision Support* 42(1), 251–266 (2006)
21. New York State Department of State Freedom of Information Laws, <http://www.dos.state.ny.us/coog/foil2.html>
22. Norman, D.: Collaborative Computing: Collaboration First, Computing Second. *Communications of the ACM*, 88–90 (1991)
23. Pardo, T., Gil-Garcia, J., Burke, G.: Sustainable Cross Boundary Information Sharing. In: *Digital Government: Advanced Research and Case Studies, and Implementation*, pp. 421–438. Springer Press, New York (2008)
24. Scholl, H.: Interoperability in e-government: More than just smart middleware. In: Proc. 38th HICSS. IEEE Press, Los Alamitos (2005)
25. Schooley, B., Horan, T.: Towards end-to-end government performance management: Case study of Interorganizational Information Integration in Emergency Medical Services (EMS). *Government Information Quarterly* 24, 755–784 (2007)
26. Siegfried, T., Grabow, B., Druke, H.: Ten factors for Success for Local Community E-government. In: Traunmüller, R. (ed.) *EGOV 2003*. LNCS, vol. 2739, pp. 452–455. Springer, Heidelberg (2003)
27. Star, S., Griesemer, J.: Institutional Ecology, Translations and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, vol. 19, pp. 387–420 (1989)
28. US Census Individual State Descriptions, <http://www.census.gov/prod/2005pubs/gc021x2.pdf>
29. West, D., Lu, J.: Comparing Technology Innovation in the Private and Public Sectors. The Brookings Institution, Washington, DC, USA (2009)

An Integrated Communication and Collaboration Platform for Distributed Scientific Workgroups

Christian Müller-Tomfelde¹, Jane Li^{1,3}, and Alex Hyatt²

¹ CSIRO ICT Centre, Australia

² CSIRO Livestock Industries, Australia

³ University of Technology, Sydney, Australia

{Christian.Mueller-Tomfelde, Jane.Li, Alex.Hyatt}@csiro.au

Abstract. We present the design, technologies and user study of an advanced collaboration platform which integrates life-size video conferencing and group interactions on a large shared workspace. The platform has been developed to support the diagnostics and research scientists in an animal health laboratory to work collaboratively across a physical containment barrier. We present the design rationale for this enhanced shared workspace which allows the sharing of a range of data and synchronous interactions on computer applications in this complex work setting. This can not be simply supported by the “board-room” type of “telepresence” technology. We describe the technical solution which has focused on the ergonomic aspect and, importantly, the integration of communication and collaboration features in the shared workspace. The platform has been under routine use and a user study has shown that these design considerations are critical for supporting the distributed scientific collaborations and may be also applicable to other scientific domains.

Keywords: Human-Work Interaction Design, Interaction with Small or Large Displays, Computer-Supported Cooperative Work.

1 Introduction

Scientific work is collaborative in nature and collaboration technologies have been increasingly seen as important to enhance the scientific collaborations. Scientific collaborations have increasingly involved research teams which are distributed. To address the problem of geographic separation in research collaborations, “collaboratories” [1], “computer-supported system that allows scientists to work with each other, facilities, and database without regard to geographical location” [2], have been emerged over the past two decades in a diversity of scientific collaboration contexts, such as physical science, biological and health sciences [3]. These “e-Science”, or “e-Research” applications came into focus with the fusion of computer and communication technologies and have demonstrated “the potential to dramatically enhance the output and productivity of researchers” [4] [5].

Designing technology to support the distributed scientific collaborations needs to move beyond developing general tools and audio-video communication

mechanisms [5]. Scientific collaboration is driven by the need to share data and to exchange and increase knowledge about the data. The collaboration and communication are different to other work practice such as business meeting and lectures with regard to the type of material and amount of data scientists work with and the data sharing process. Various sources of data and information need to be “at hand” to be reviewed and interpreted by different experts at the same time rather than enabling one-to-many, lecture-style presentations.

Today’s video conferencing technologies are highly developed in terms of audio and video quality and aiming for reproducing face-to-face situations, such as the specialized board meeting room for telepresence [6]. However these solutions may not be suitable to support scientific collaborations since they have limited support for sharing and working with electronic documents - the shared data space is either small or is compressed in its visual precision.

Recent research in collaboration technology have explored shared workspaces and interaction techniques that enable data sharing within multi-display environments, such as WeSpace [7] and Impromptu [8]. These works address the co-located team meetings, including scientific collaboration situation. Inspired by these, other work, such as iBIS [9], has demonstrated a multi-display and coherent physical environment to support various interactions including distributed collaborations.

Developing technologies to support the real-world distributed scientific has been considered as necessary to understand and support the communication aspect and the dynamic of information exchange [10]. A successful collaboratory must respect users’ existing communication and work mechanism. An understanding of how collaboration work should be done prior to the design and users’ involvement in the design can not only allow rapid development cycle [11] but also influence the adoption of technologies and long-term outcomes which are the criteria for evaluating collaboratories [12].

Furthermore, Hollan and Stornetto [13] pointed out, that communication and collaboration tools should provide solutions “which are not ideally met in the medium of physical proximity, and evolving mechanisms which leverage the strengths of the new medium to meet those needs”. In other words, rather than mimicking and supporting existing co-located collaboration practices for distributed scenarios, the new solutions should aim for enabling distributed teams to collaborate “beyond being there”. Distributed scientific collaboration and communication needs exactly this kind of tools and solutions, not only to work as good as being co-located but also to increase productivity and creativity.

In this paper, we present the design process, the technical solution and the early user experience of a collaboration platform which integrates life-size video conferencing and group interactions on a large shared workspace to support distributed scientific collaborations. Our attention in this paper is directed towards supporting scientific collaboration and communication in the unique environment of the Australian Animal Health Laboratory (AAHL) which has high level of physical containment (PC) environment. Based on the foundation of prior work in collaboration technologies, we are motivated to contribute to the research field with a case of applying an integrated communication and collaboration platform in a scientific collaboration work environment.

2 Background

Australian Animal Health Laboratory (AAHL) plays an important role in animal disease diagnosis, research and policy advice in Australia. Through its ongoing research programs, AAHL is able to develop the most sensitive, accurate and timely diagnostic tests, which are critical to the success of any eradication campaign in the event of a disease outbreak. AAHL also undertakes research to develop new diagnostic tests, vaccines and treatments for both exotic and endemic animal diseases.

AAHL is a “secure” scientific laboratory which has high biocontainment facility allows safely handle a range of animal species to physical containment level three (PC3) and the highest level, PC4. These levels reflect the risks involved handling biological substances. However, this facility also poses a challenge in terms of effective, rapid communication and data sharing as when leaving the containment areas staff needs to have thorough shower and equipment needs to follow strict rules of a decontamination procedure.

Real-time interactions between scientists and sharing scientific data and resources across the barrier are critical for AAHL to provide diagnostics and research services, particularly in the context of the time pressures of an emergency disease outbreak. The effort required to go through the containment barrier introduces communication difficulties between scientists working at the different physical areas. Different groups of research and diagnostics scientists need to work together and have various data resources to work with on a regular basis, such as data from high performance microscopes, shared image data bases and electronic notebooks. Email and telephone were the common communication tools. Quite often staff had to spend time going through the containment barrier to have face-to-face meetings, or meetings need to be scheduled not only with respect to availability of staff but also to the areas they are at the time. The collaborations in this environment cannot be clearly characterized as traditional “same place” or “different place” collaboration as defined by the time-space matrix [14]. Rather the collaborations across the containment barrier need be considered as in-between: collocated and remote, relating to the context of the team working at areas of different containment level but within one social organization and one physical building.

3 Design Process

Our goal was to develop a platform to support scientists to effectively communicate and share information across the containment barrier. The one-year design process included a field study to understand the practice, scenario-based use case analysis, iterative design, mock-up, user testing and deploying.

3.1 Understanding the Practice

A field study was conducted at the beginning of the project. Based on eleven semi-structured interviews with different work groups, one focus group meetings with twenty staff and four site visits, we tried to build a picture of users’ work practice, particularly collaborations. We had regular meetings with users in the project period and invited

users to review and test the design before the platform was delivered. Users' continuous involvement in the design was part of the iterative design process. These understandings led to the development of three central user scenarios which helped to discuss the features of the platform between the design team and the users [15].

The diagnostics and research work in AAHL are not confined to one specific work group and often involve both staff working in the containment areas and staff working in the general office area. For example there are meetings each week between the diagnostics management team (such as the veterinary officers) who work in the general office area and scientists who work in the containment area. The veterinary officers are responsible for delivering timely technical report to state or territory animal disease control authorities. It is necessary for the veterinary officers to analyse and interpret the results together with the diagnostics scientists and microscope scientists who conduct diagnostics tests inside the PC3 area. Similarly the science research work often involves research scientists outside the PC3 area, diagnostics scientists and microscopy scientists in the PC3 areas and requires efficient data sharing between them.

During our site visits, we observed the constraints on the collaboration practice in this laboratory. The containment barrier not only separates team members spatially, but also makes their workflow difficult to organise, for example to join a meeting outside, scientists inside have to "shower out" and therefore need to schedule their activities to avoid unnecessary pauses in their work and to avoid having multiple showers a day. User scenarios have been carefully identified after discussions with the key representative staff of AAHL. The proposed scenarios address the communication and data sharing issues caused by the specific barrier, particularly the collaborative group meeting in which a group of scientists inside the containment area to work with a group scientist in the general office area. There are different types of meeting situations, ranging from conversational meetings to data-centred meetings and it is important to support both of these requirements. The technical solution (see Figure 1), a shared workspace inside the containment area and a shared workspace in the general office area, was designed and developed to mitigate the impact of moving between the two areas, to facilitate communication and to support the free flow of information and application sharing.

3.2 Iterative Design and Testing

At the beginning of the design process, a technical demonstration based on our previous work [9] was set up at our premise and tested by two design representatives from AAHL. Based on their hands-on experience, they were able to quickly contribute to requirement specification which helped us to generate an initial technical solution. During the design phase, different layouts, size and position of the equipment were carefully evaluated by the technical design team using existing hardware components and digital mock-up and animations. Devices were carefully chosen and tested for sufficient quality. After these steps, a refined design of the "shared workspace" was reviewed by three design representatives from AAHL and two experts from the national biosecurity authority. This feedback of the expert users were taken into account before the platform was finalised, developed and commissioned. After the introduction of the platform and training to the AAHL staff, a user study was conducted to capture the early adoption feedback from users (see section 5).



Fig. 1. The integrated communication and collaboration platform at Australian Animal Health Laboratory (AAHL)

4 The Integrated Communication and Collaboration Platform

In this section we describe in more details the technical aspects of the platform and explain the design rationale as an outcome of the interactive process. Our considerations focused not only on the technical solutions of communication and application sharing, but also the appropriate integration of these features and the associated consideration in the design of the physical space of the platform. The communication part of the platform delivers a dual HD video link between two platforms including high quality, echo-cancelled audio communication. In addition the 8.3 Million Pixels computer workspace can be shared and displayed across the sites.

The hardware components of the platform itself are available off-the-shelf and software is mostly open source. However the close integration and flexible control of functionality of the platform are unique and to the best knowledge of the authors not achieved elsewhere, either by available products nor by related prototype from the research community. In the following sections, we will describe the physical design of the platform, and the integrated communication and collaboration technologies, as well as central user interface to the platform.

4.1 Physical Design of the Platform

We designed the platform in accordance with standard work place ergonomics. The table in front of the display is long enough to accommodate up to 4 users (see Figure 1). The table height of 72 cm and depth of 75 cm are those of standard office tables. The display panels with a resolution of 52 pixels per inch (ppi) suggest an optimal viewing distance of 1.66 m with respect to normal visual point acuity of

users. Therefore, the complete display space lies within the limits of the users' visibility for all seating positions at the table, approx. 1.25 m away from the display. To avoid uncomfortable viewing angles to the top displays we decided to lower the all display as much as possible and to tilt the top displays (see Figure 2). We provided a keyboard and three mouse input devices to allow standard interaction with the computer. Two video cameras are positioned to capture all four users sitting at the table. Therefore, readjusting the cameras' orientations or field of views are not required, sitting at the table implicitly results in the fact that the users are "in the picture" (similar to systems described in [6]). In that way the overall affordance of the setting comprises the affordances of media spaces (as described in [16]) with those of the physical office environment including the chairs and the table. The table also allows to be brought into the meetings all sorts of artefacts such as pen and paper, laptops, or coffee mugs, and to place them on it.

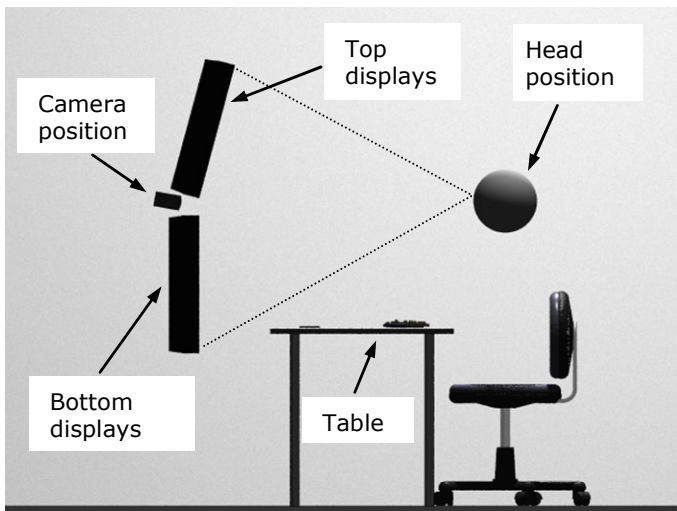


Fig. 2. The orthographic side view of the physical setting of the platform as shown in Figure 1. The bottom displays are lowered below the table surface to allow the top displays to be in a comfortable viewing position for the users.

The platform consists of four 42 inch diagonal Liquid Crystal Displays (LCD) units each with a pixel space of 1920x1080. Hence, the total display space is 3840x2160 pixels (approx. 8.3 Million Pixels) at a physical resolution of 52 ppi. All four displays are driven by one computer. The bottom two displays are mounted vertically on a frame structure and lowered to the height of 60 cm above the ground to allow the top displays to be lowered as well for comfortable viewing. The top displays are also mounted on the frame structure and tilted to further support a comfortable viewing by users sitting at the table. A gap between the edge of the table and the bottom displays guarantees that users sitting at the table can easily oversee the whole display space (see Figure 2). Digital models of the platform components have allowed

us to assess design variants during the design process and helped us to develop the platform to the current state.

4.2 Video Conferencing

We use two off-the-shelf videoconferencing units per site to deliver the video and audio link between two sites. In a 10 cm gap between the two rows of displays we mounted two High Definition (HD) cameras that capture the users sitting in front of the displays. The camera positions provide good eye contact perception when the remote site is displayed on the bottom displays. The two HD cameras create a near life-size video image on the remote displays and are arranged to maintain spatial continuity of the images. Two microphones positioned on the table and connected to one conferencing unit capture the users' voices on each site.

There are three remote video display modes, "full view", "picture-in-picture", and "hide" mode (see Figure 3). In the "full" mode, the entire two bottom displays are used to show the remote site. In "picture-in-picture" mode, the video of the remote site is reduced to 25% (compared to full mode) and positioned on the top displays, left and right from the middle at the lower edge of the tops displays. Finally, in the "hide" mode no remote video image is displayed, however audio communication remains always on. These modes allow users to switch quickly between different degrees of video media richness depending on the purpose of the meeting and particular interaction situations during the meeting.

4.3 Application Sharing

For the shared collaboration workspace we use a VNC server running on a dedicated server computer [17]. The client computers on each side of the collaboration platform are connected via a Gigabit network connection to this server. The clients connect to the server running a VNC viewer program [18] in full screen mode. The server is running without monitors and is configured to automatically provide the shared server facilities after restart. In that sense the client computers act like kiosks, or thin clients which have the only task to display the server computer's desktop, as typical for desktop virtualization. However, in the presented setting the size of the virtualised and shared desktop of 8.3 Million Pixels is roughly four times that of a typical personal desktop computer.

To comply with the laboratory's regulations and policies with respect to accessing computer resources we set up the application sharing in the following way. The only available applications on the shared server workspace are: a) Remote access programs to desktop computer, dedicated instrument computer, and terminal server computers, or b) a general purpose web browser. Allowing only these applications to be launched on the platform's shared workspace is compliant with the computer access policies of the laboratory. The access to intranet databases or other web-based services complies with staff's existing data access security configurations. This approach provides high flexibility of visualizing multiple desktops of remote computers and applications and the same time guarantees that no unauthorized access happens. No classified data is left on the platform's shared workspace server and client computers after the meeting.

4.4 User Interface Integration

In order to provide an easy use of the platform we provided a control application to access the essential functions of the platform, such as turning on or off the video connection and the workspace sharing. One design goal of the integration was to eliminate the use of the up to 6 remote controls of the two video conferencing units and the four displays and to integrate of all platform functions at the user interface level to avoid cognitive overload (see Figure 3). Another design goal was to provide users various ways to control the platform, either through a central Graphical User Interface application with all functions or through the multi-media buttons of the computer keyboard for functions such as displays on/off or changing the remote video display mode.

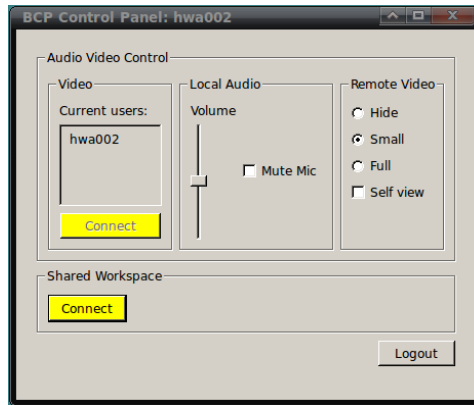


Fig. 3. The central user interface to control the platform. The interface provides all necessary functions and replaces all device remote controls.

5 Early User Experience

At the time of writing, the two platforms have been used at AAHL by various work groups for nearly half a year. Some of the groups have used the platform on a routine basis, other in a more ad hoc way. We conducted a user study to understand the usage of the platform in AAHL after it was used for two months. Based on a combination of interview, questionnaire and observation methods, the user study tried to capture the users' early experience with the platform.

21 staff from five work groups took part in the study. This was around 50% of the total number of staff in the five work groups who were regular or potential platform users. These participants were identified as staff who have used the platform. 5 semi-structured interviews were conducted. One collaboration meeting was observed with audio-video recording of the meeting.

The results of our questionnaire and interviews have shown that study participants clearly articulated that the platform was a useful tool to enhance the communication and to enable efficient sharing information across the biocontainment barrier.

The features of the ability of sharing and interacting with multiple data resources and high quality audio and video have been highlighted by the study participants.

The result of the questionnaires reveals that usually the number of the meeting participants were 3 to 4 at each end and the meetings lasted around 50 minutes. Most of the meetings were planned regular group meetings while some meetings were ad hoc and for special purposes. The participants rated the platform as ‘easy to use’ (2.4) based on a 5-point rating scale (1: very easy, 3: medium and 5; very hard). When asked about the usefulness and helpfulness of the platform compared to face-to-face meetings the average user reported “the same” (2.9) on 5-point rating scale (1: much worse, 3: the same, and 5: much better). Some participants commented that they felt that the system was better than fact-to-face because of the real-time access to and the availability to share and collaborate on all relevant data during the meetings. Although some participants felt that it can “never replace fact-to-face meetings”, they valued the platform as “from a cost perspective, it helps”. Our findings showed that 40.9% of the participants found that the choice of display mode depended on the situation during the meeting. This reflects the design requirement of supporting different types of working interactions (e.g. conversational meetings or data-centred meetings) and a flexible configuration of people view on the displays.

In the interviews and the questionnaires, study participants gave comprehensive feedback to the questions of what they “liked and not liked” about the platform. This feedback captured their understanding of the platform and highlighted areas for improvement. It was pointed out as a drawback that the current platform only allows one person to control the mouse at a time. Being able to work on documents in a private workspace before sharing them on the shared workspace was also mentioned as an area for improvement since there might be some sensitive documents or applications which participants may not want to share without preparations. Study participants have expressed strong interests in extending the current platform to support other collaboration scenarios, such as the collaborations with research partners outside AAHL. We also found that some users tend to use the platform for co-located meetings and make use of the large display space to share multiple personal desktops for discussions.

6 Discussion and Future Work

Early user experience with the platform has confirmed our design rationale of the platform. The platform meets the specific requirements of the distributed scientific collaborations in this laboratory. Users received well the benefits of the integrated platform and conceptualised the potential of a real-time information sharing in supporting their collaborative work. The feature of sharing and interacting with a broad range of data resources in large displays support the fundamental focus of scientific collaborations, particularly group collaborations. Together with the ability to quickly reconfigure the display space, the platform may support user experience that can be characterised as beyond being there. In other words, the platform not merely brings together scientists separated by the containment barrier, but also provides at the same time a flexible data sharing and communication environment that can be adjusted to different communication and collaboration needs during the scientific meetings.

The fact that users also use the platform for co-located meetings without making use of the videoconferencing capabilities may also provide evidence for the usefulness of the platform and worth to have further investigations.

We will conduct field observations and survey over a period of several months in a longitudinal study to have in-depth understanding of the platform usage and user experience. As part of the continuous iterative design process, the areas for improvement identified from the user study will help us to refine the platform. For example, we will extend the current use scenario by providing laptop computers at each site for managing private workspaces that can be shared to the platform in an ad hoc way. Also, we hope that the integrated collaboration and communication platform we have designed and developed can be extended to support the research and operational collaborations between AAHL and a number of other organizations.

7 Conclusion

We described the design rationale and technical details of an integrated platform for scientific collaboration and communication. In the design process we engaged the potential user group to provide an effective technical solution to fit into the particular scientific laboratory environment. Early user feedback supports our design goal and we expect to confirm this trend further in a longitudinal study. Our design solution differs from other systems in the fact that the platform provides users with a large shared workspace for real-time collaboration in combination with high quality videoconferencing. The integrated platform bears the potential to be applicable to other domains, where distributed communication and collaboration on scientific data is the key of the application.

Acknowledgments. This work is part of a National eResearch Architecture Taskforce (NeAT) project, supported by the Australian National Data Service (ANDS) through the Education Investment Fund (EIF) Super Science Initiative, and the Australian Research Collaboration Service (ARCS) through the National Collaborative Research Infrastructure Strategy Program. We thank other members of the technical team from CSIRO ICT Centre for their contributions and Martyn Jeggo from AAHL, John Taylor from CSIRO CMIS for their support.

References

1. Wulf, W.A.: The national collaboratory: A white paper. In: Appendix A in Toward a National Collaboratory, 1 (1989); unpublished report of a National Science Foundation invitational workshop held at Rockefeller University
2. Finholt, T., Olson, G.: From laboratories to collaboratories: A new organizational form for scientific collaboration. *Psychological Science* (1), 28–36 (1997)
3. Olson, G.M., Zimmerman, A., Bos, N.: *Scientific Collaboration on the Internet*. MIT Press, Cambridge (November 2008)
4. Chen, M.: Leveraging the asymmetric sensitivity of eye contact for videoconference. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2002)*, pp. 49–56. ACM, New York (2002)

5. Chin Jr., G., Lansing, C.S.: Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory. In: Proceedings of the of the ACM Conference on Computer Supported Cooperative Work, CSCW 2004, pp. 409–418. ACM, New York (2004)
6. Gorzynski, M., Derocher, M., Mitchell, A.S.: The Halo B2B Studio. In: Harrison, S. (ed.) Media Space 20 + Years of Mediated Life. Computer Supported Cooperative Work, pp. 357–368. Springer, London (2009)
7. Jiang, H., Wigdor, D., Forlines, C., Shen, C.: System design for the wespace: Linking personal devices to a table-centered multi-user, multi-surface environment. In: Tabletop, pp. 97–104. IEEE, Los Alamitos (2008)
8. Biehl, J.T., Baker, W.T., Bailey, B.P., Tan, D.S., Inkpen, K.M., Czerwinski, M.: Impromptu: a new interaction framework for supporting collaboration in multiple display environments and its field evaluation for co-located software development. In: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (CHI 2008), pp. 939–948. ACM, New York (2008)
9. Broughton, M., Paay, J., Kjeldskov, J., O’Hara, K., Li, J., Phillips, M., Rittenbruch, M.: Being here: designing for distributed hands-on collaboration in blended interaction spaces. In: Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group (OZCHI 2009), pp. 73–80. ACM, New York (2009)
10. Henline, P.: Eight collaboratory summaries. *Interactions*, 66–72 (May 1998)
11. Finholt, T.A.: Evaluation of electronic work: Research on collaboratories at the University of Michigan. *SIGOIS Bull.*, 49–51 (December 1995)
12. Sonnenwald, D.H.: Expectations for a scientific collaboratory: a case study. In: Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work (GROUP 2003), pp. 68–74. ACM, New York (2003)
13. Hollan, J., Stornetta, S.: Beyond being there. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1992), pp. 119–125. ACM, New York (1992)
14. Baecker, R.M., Grudin, J., Buxton, W., Greenberg, S.: *Readings in Human-Computer Interaction: Toward the Year 2000*. Morgan Kaufmann, San Francisco (January 1995)
15. Li, J., Müller-Tomfelde, C., Hyatt, A.: Supporting collaborations across a biocontainment barrier. In: Proceedings of the 2010 Conference of the Computer-Human Interaction Special Interest Group (CHISIG) of Australia on Computer-Human Interaction (OZCHI 2010), Brisbane, Australia, pp. 320–323 (November 2010)
16. Gaver, W.W.: The affordances of media spaces for collaboration. In: Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work (CSCW 1992), pp. 17–24. ACM, New York (1992)
17. TightVNC: VNC-Compatible Free Remote Control / Remote Desktop Software (2010), <http://www.tightvnc.com/> (cited 23.01.2010)
18. RealVNC - VNC remote control software (2010), <http://www.realvnc.com> (cited 23.01.2010)

IdeaTracker: An Interactive Visualization Supporting Collaboration and Consensus Building in Online Interface Design Discussions

Roshanak Zilouchian Moghaddam, Brian P. Bailey, and Christina Poon

University of Illinois at Urbana,
Department of Computer Science
{rzilouc2,bpbailey,cpoon2}@illinois.edu

Abstract. With the rapid growth of open source and other geographically distributed software projects, more interface design discussions are occurring online. Participation in such discussions typically occurs via issue management systems or similar interactive discussion forums. While such systems have a low learning curve, they do not support key elements of design discussion such as comparing alternatives, maintaining awareness of the arguments for and against the alternatives, or building consensus. To better understand these and other challenges, we conducted a study of online interface design discussion. The study consisted of analyzing a large corpus of online discussion content and conducting interviews with designer and developer participants. We discuss the findings of our study and use them to motivate the implementation of an interactive visualization tool - IdeaTracker. The tool offers explicit support for tracking and comparing ideas and gaining an abstract summary of the overall discussion as well as specific alternatives. It also provides a voting system to support consensus building. The tool extracts and visualizes useful information from the discussions that would otherwise be hidden but without interfering with the current method of participation. Our tool is compatible with the issue management system of one open source project but can be extended for others. Initial user feedback is positive and confirms the need for an alternative visual representation of interface design discussions online.

Keywords: Design, open source software, user interface, visualization.

1 Introduction

The User Interface accounts for a large part of the design and development effort in the production process of any interactive software system [18]. With the number of geographically distributed software projects and, in particular, open source software projects growing [10], more interface design and development discussions are unfolding online. For example, in open source projects, designers typically engage in interface design discussions through the project's issue management system. These systems offer interactive Web-based forums for discussing design issues. These forums are similar to

other types of Web-based newsgroups and social media where participants discuss a variety of topics from food and books to religion and politics.

However, interface design discussions have many unique characteristics that are ignored by the current metaphor of interactive forums. For example, common practices that are paramount to producing effective interface designs such as generating, discussing, and comparing multiple ideas [9, 26] are not specifically supported. This problem is exacerbated due to the diverse levels of expertise and viewpoints of the discussion participants [17].

A few studies have analyzed design discussions in open source software. For example, Barcellini et al. studied software discussions occurring in mailing lists with the goal of extracting design relevant information [7]. Unlike their study, our work focuses on analyzing and enhancing design discussions relating to the *interface* of a software project. Twidale et al. analyzed usability reports from Bugzilla [25]. They recommended several improvements such as enhancing the classification of usability bugs and adding explicit representations of design arguments to the reports. However, these enhancements were never implemented and their study considered only usability bug reports rather than, e.g., proposals for new interface features.

Various tools have also been developed to support distributed interface design. For instance, Designers' Outpost [11], Synergy [15] and the Envisionment and Discovery Collaboratory [4] are examples of collaborative design tools that can be used for co-located or distributed team situations. In addition, a myriad of tools have been developed to support the creation of design artifacts in distributed settings [1-3]. Though these types of tools can facilitate the early stages of interface design, our work is original in that we are targeting improving online *discussions* of interface design issues. Though our current focus is on open source projects due to the public availability of discussion data, we believe our results are applicable to any situation where Web-based interactive forums are being used as the primary means for discussing interface design issues.

To better understand designers' challenges participating in online design discussions, we studied a large corpus of interface design discussions from two popular open source projects: Ubuntu and Drupal. In our study we analyzed the content of 1560 messages collected from thirty discussion threads from the Ubuntu and Drupal issue management systems. Figure 1 shows an example of a design discussion extracted from the Drupal issue management system. Our analysis included counting the posts, alternatives, and participants per discussion as well as the fluidity of the topic flow. Our data analysis was complemented by interviews with UI designers (N=6) and developers (N=6) participating in the design discussions. The interviews helped interpret the results from the content analysis and uncover the current nature of participation, idea generation, and the consensus building process.

From data analysis and interviews, we found that designers are struggling to track the design alternatives and the arguments for and against each alternative, want to maintain better awareness of others' preferences, and prefer that the interface design and development activities remain tightly integrated. We have incorporated these and other findings into the design of a new interactive visualization tool called IdeaTracker for managing online interface design discussions. The visualization highlights the proposed alternatives in a visual timeline, allows participants to filter and review messages that refer to each alternative, and provides an abstract visual summary of each message to

indicate length and affective tone of each message. IdeaTracker also helps with decision making by enabling designers to compare selected subsets of alternatives and supports consensus building by implementing a voting system. We believe that using IdeaTracker can result in more effective design discussions online, which can lead to higher quality interfaces. Initial user feedback is positive and confirms the need for an alternative visual representation of interface design discussions online.

Our work makes two contributions. First, we identify key challenges participants face when engaging in interface design discussion online, which include tracking and comparing alternatives, maintaining awareness of community opinion, and building consensus around specific alternatives. Second, we demonstrate how these challenges can be addressed through the implementation of a novel interactive visualization tool. The tool extracts and visualizes useful information from the discussions that would otherwise be hidden but without interfering with the current method of participation.

Issues

Vertical tabs - no visual connection between the active tab and its content

Posted by [Manuel Garcia](#) on August 29, 2010 at 7:19pm

Description

In a talk in the Drupalcon CPH, it was mentioned that currently the relation between a selected tab and the content it displays is not clear at all, and that we should perhaps work on this and figure out a better way to make this visual connection. I attach a current (verticaltabs-seven-before.png) and proposed (verticaltabs-seven-after.png) way to make the user aware of what tab he is using.

Attachment	Size	Status	Test result	Operations
verticaltabs-seven-before.png	16.95 KB	Ignored	None	None
verticaltabs-seven-after.png	23.13 KB	Ignored	None	None
verticaltabs-seven.patch	1.81 KB	Idle	PASSED: [[SimpleTest]]: [MySQL] 23,316 pass(es).	View details

Comments

#1 Posted by [Manuel Garcia](#) on August 29, 2010 at 7:20pm
Unsure whether this is actually a bug, but well, feel free to slap me with a large trout! :x

#2 Posted by [Jeff Burnz](#) on September 1, 2010 at 9:43am
I think it is a bug - if forces the user to continuously adjust the scroll - for example on every node going from Menu tab to others changes the height drastically (default install). For daily usage this is quickly going to be a huge PITA for content editors.

#3 Posted by [Manuel Garcia](#) on September 1, 2010 at 12:57pm
Although I see what you are talking about Jeff, I think that's not related to this issue. Please take a look at the screenshots to see what I'm talking about, I don't touch the fact that if you open a very long form on a tab (which is discouraged as bad practice afk) you'd have to scroll down, then back up to change tab.

First Alternative

Fig. 1. A sample interface design discussion thread in Drupal. The interface design problem discussed in this thread is the lack of visual connection between a selected tab and its content. An alternative has been proposed along with the problem description. In the first few comments, participants are discussing the scope of this problem and whether it is a valid issue.

2 Related Work

We describe prior work for representing online content in different domains and how our study builds upon and extends that work for online interface design discussion. Also, we describe open source and the usability enhancement movement within open source and how our work can aid this movement.

2.1 Distributed Collaborative User Interface Design

A myriad of tools have been developed to support creation of design artifacts in distributed settings. For instance Axure [1], Pencil [3], and Balsamiq [2] are part of a large class of tools that have been developed to manage online collaborative creation of interface mockups. These tools assume that collaborators utilize existing tools such

as interactive discussion forums or synchronous messaging systems for discussing the designs. Other tools have been developed to support remote collaboration in UI design. For instance, Distributed Designers' Outpost is a collaborative web site design tool that captures the history of an evolving information architecture [11]. Synergy is another collaborative design environment that supports remote collaboration during the early stages of the design process [15]. However, there has been little research into interfaces that specifically support the unique nature of *online design discussions*. We have taken a first step in this direction by first understanding specific challenges of such discussions and showing how they can be addressed in a novel visualization.

2.2 Visual Representation of Online Content

A visual representation of online content expressed in narrative form can foster sense making of the information. Such visual representations have been created for visualizing content in myriad online communities. For instance, Opinion Space is an online interface designed for collecting and visualizing users' opinions from comments in online news articles, blogs, videos, and product reviews [12]. Shared Space is another tool that analyzes students' chat messages to visualize discussion and agreement during online discussions [14]. The Conversation Map system creates a graphical interface by analyzing the text of archived newsgroup messages. The interface can be utilized to read and search all of the messages in the archive [22].

All of these tools were developed to represent a specific type of discussion content in an online community: Opinion Space represents opinions, Shared Space represents chat messages, and Conversation Map shows newsgroup posts. Similarly, interface design discussions have certain elements that, if incorporated into the design of a visual representation, can foster participation and improve the quality of discussions. Inspired by this previous work, this paper explores the design of an interactive visualization tool that supports interface design discussions, capturing the unique characteristics of this content.

2.3 User Interface Design in Open Source

A usability movement has been ongoing in the Open Source community [5, 8, 13, 16, 19, 23, 24]. Corporations were one of the first contributors to this movement by providing interface guidelines and recruiting usability experts to work on open source projects [5, 8, 19]. The open source community itself helped this movement by promoting new techniques such as "Design-by-blog" where personal blogs are used for design discussions [20]. In many projects, screenshots can now be posted in mailing lists and issue management systems which facilitate design discussion [25]. Also, usability experts have started joining interface design discussions in open source through blogs, issue management systems, and wikis [5, 27]. As more designers, usability experts, and developers engage in discussions of interface design issues for open source, there is a growing need for studies to characterize their participation challenges and identify feasible ways to address them.

As a starting point, several studies have been conducted to identify the challenges of improving the user interfaces of open source projects [5, 8, 13, 16, 19, 23, 24]. Some solutions propose to resolve the challenges by funding usability experts to work

on open source projects, adapting automatic approaches for evaluating interfaces, encouraging HCI students to join open source, and providing a social and technological infrastructure for usability experts [19]. Bach et al. expanded this last solution by arguing for a separate design space for usability experts at the same level as developers in the project's information (and social) architecture [5].

Our work shares the long-term goal of improving the user interface of open source projects, but emphasizes enhancing the discussions of interface design issues, thereby leading to more effective participation. In contrast to Bach et al., our approach is to provide a more effective visual representation of the existing design discussions, which involve both designers and developers, rather than advocate for a separate discussion space targeted only for designers.

Finally, in relation to a prior study of online interface design discussion for open source [25], our work is original in that it analyzes a broader sample of design discussions, complements this analysis with results from stakeholder interviews, and implements the findings within an interactive visualization tool.

3 A Study of Online Interface Design Discussion

We studied the evolution of interface design discussions in two well-known open source projects: Ubuntu and Drupal. Ubuntu is a popular Linux distribution and Drupal is a widely used content management system. We chose these two projects because they have an active interface design team, the design discussions are publicly accessible, and they contain a considerable amount of lively design discussion.

Within both of these projects participants post interface design problems, solution proposals, and related comments to the respective issue management system. These issue management systems are Web-based interactive forums dedicated to discussing software related issues including interface design issues. These forums are similar to the interactive forums used for discussing topics of interest in many other online communities or social media sites.

As shown in Figure 1, each interface design discussion (a new thread) starts when a participant posts the description of a design problem. Other participants can then contribute to the discussion by proposing design alternatives, arguing for or against the proposed alternatives, attaching an implementation of an alternative (called a *patch*), reporting the results of a patch review, or raising other concerns such as clarifying the scope of the problem or how it relates to other ongoing design efforts.

Generating multiple alternatives and having an engaged discussion are critical for producing an effective solution to the design problem and having confidence in it. Designers also need mechanisms to share opinions and increase awareness of others' perspectives to facilitate consensus. This is difficult even during face-to-face design meetings, yet the issue management systems used for interface design discussions only support text postings along with the option of attaching an image. They do not have explicit support for key elements of the discussion itself such as tracking which ideas are favored, arguments for or against the ideas, or building consensus around specific alternatives. Our study therefore centered on answering these questions:

1. How many ideas are typically shared in online design discussions? How extensive is the debate? How do designers maintain awareness of the ideas?
2. What techniques are used to promote consensus around specific alternatives and how effective are these techniques?
3. What are the strengths and weaknesses of the issue management systems used for discussing and managing ideas?
4. What are other challenges that designers encounter participating in online design discussions and what strategies are used for addressing them?

To answer these questions, we analyzed content from online interface design discussions and conducted interviews with participants. For the quantitative analysis, we examined 1560 messages that spanned thirty discussion threads. Fifteen of the threads came from a pool of 300 in Ubuntu (average number of messages in each thread=35.4 $sd=12.9$). The other fifteen come from a pool of 500 threads in Drupal (average number of messages in each thread=68.6 $sd=23.0$). In both cases, we selected the fifteen threads from the most active threads. From inspecting a sample of the threads; we defined two characteristics for ‘activity;’ (i) the number of images, where we considered an image to be a proxy for a proposed design alternative and (ii) the number of messages in that discussion. We rank ordered the pool of discussion threads based on these criteria and selected fifteen threads from the top fifty threads in each project.

We then analyzed the content of the messages. First, we divided the message into a smaller set of topical chunks. A new chunk was created when there was a transition from one topic to another. Then, we manually coded the chunks into eight categories: Issue, Alternative, Criterion, Clarification, Project Management, Implementation, Digression, and Other. We adapted these categories from the coding scheme developed by Olsen et al. [21] for capturing and analyzing the core elements of design discussions in collocated settings. Though the original coding schema had eleven categories; our adaptation only used seven. When testing the schema and resolving inconsistencies (based on five discussion threads from each project), the evaluators agreed that the other four categories were not applicable. An implementation category was added to capture the technical messages in the discussions.

The data analysis was complemented by a set of semi-structured interviews with twelve participants from both Ubuntu and Drupal projects. We interviewed six active designers, three from Drupal with an average of 7 years of experience ($sd = 1$) and three from Ubuntu with an average of 7 years of experience ($sd = 3.26$). We will refer to Drupal designers and Ubuntu designers as DD# and DU# accordingly. We also interviewed six developers participating in resolving UI design and usability issues: three from Drupal with an average of 6.5 years of experience ($sd = 3.51$) and three from Ubuntu with an average of 7.5 years of experience ($sd = 4.5$).

Each interview lasted about an hour and was conducted via phone and instant messaging. The latter was used to share Web links, images, and other data artifacts during the interview. The subjects were compensated with a \$30 Amazon gift card.

The interview questions reflected the main research questions of our study and were informed by prior work (e.g. [5]). For example, the questions included: what are the main challenges in discussing design ideas online? How do you maintain awareness of the ideas? What techniques are used to promote consensus? What are the strengths and weaknesses of the issue management systems for discussing ideas?

4 Challenges and Implications

We discuss key challenges that designers encounter while participating in interface design discussions online and the methods used for addressing those challenges. Although our study focused on open source projects, the challenges identified were mostly due to the intersection of the interaction design of the forums and the topic of the discussion. As a result, we believe these observations apply to any interactive forum which serves as the primary means for discussing interface design issues.

4.1 Designers Struggle to Track Design Alternatives

A total of 299 alternatives (ideas) were submitted in the 1560 messages analyzed. On average, nine alternatives were proposed in each thread of discussion ($sd=5.88$, $max=27$, $min=0$), indicating that multiple alternatives were welcomed and considered for each design problem. Of all the alternatives proposed, 63% were described solely in narrative form, 13% included only a patch (a file that updates the implementation to provide a working preview of the idea), and 18% included only a screenshot of a proposed solution. Figure 2 shows a distribution of alternatives for each combination of modality. Although we chose part of our data set from the discussions with the most images, more than half of the alternatives were presented in narrative form. This is partly due to the lower cost of expressing alternatives in narrative form. Other reasons for having fewer visual representations may be the text-based structure of the discussion threads and lack of emphasis on the importance of visually demonstrating alternatives by the discussion interface. However, all of the designers were aware of the importance of screenshots and other visual representations in explaining a design alternative. They mentioned that alternatives with visual representations have a better chance of receiving comments from community members. For instance DU2 said: “...wireframes or other visual stuff legitimize ideas. Because they’re very memorable. [...] People comment on images more, or codes more than paragraphs.”

One way to improve the discussion interface is to better emphasize the importance of alternatives and the significance of providing visual representations for them by adopting a more visual structure.

Confirming the importance of alternatives, our quantitative analysis revealed that 48% of the conversation is spent discussing alternatives (Figure 3). Designers also mentioned alternatives as a vital piece of information when contributing to a design discussion. As DD1 said: “[When participating in a new discussion] I would want to know, in its current state, what is the exact problem the issue is trying to deal with, what are the proposed solutions so far, and what direction have people been taking on each of the proposed solutions.”

However the current systems do not support tracking the proposed alternatives. As DU3 said: “They (bug reports and mailing lists) aren’t good for keeping track of all of the ideas. [...] There usually isn’t anybody keeping track of these are all the possible options and these are some [discussions] about each option.” Today, the only way a participant can maintain awareness of the proposed alternatives is to rely on their memory, maintain notes in a separate tool, or scan the entire discussion thread to review the alternatives and rationale for and against each of them. The first method is unreliable as human memory is fragile and has a limited capacity [6]. The latter two methods are cumbersome and do not promote shared awareness among participants.

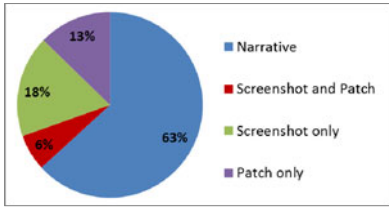


Fig. 2. Distribution of alternatives over each combination of modality, aggregated across all messages (N = 1560)

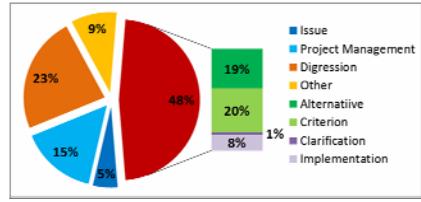


Fig. 3. Distribution of each coding category. Summing the categories for Alternative, Criterion, Implementation, and Clarification indicates that 48% of all discussion was devoted to alternatives.

One of the consequences of using an interactive forum for managing a design discussion is that many of the proposed alternatives get buried in the midst of the discussion. As DU4 said: *“Some ideas that people have are actually really good, but then they kind of get lost in the thread...”* This means that not all alternatives will be evaluated thoroughly or may be forgotten during a lengthy discussion.

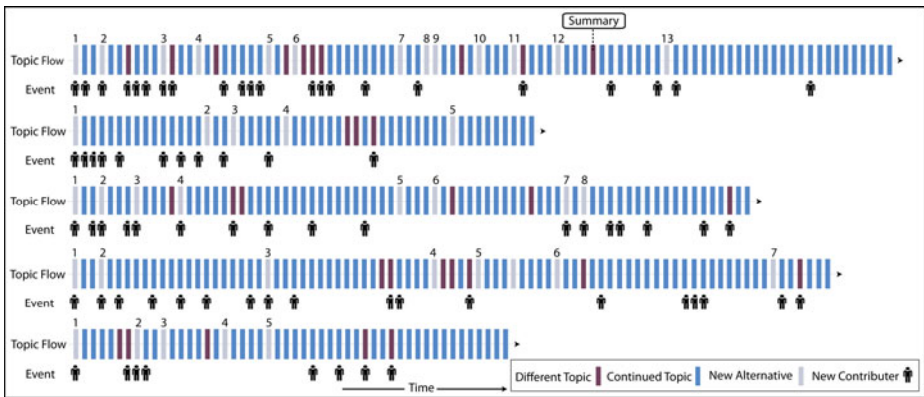


Fig. 4. An event timeline for five complete threads of interface design discussion from the Drupal project. The timeline shows the fluidity of topics along with visual indicators of when new alternatives were posted and when new community members joined the discussion thread.

Figure 4 demonstrates some of these challenges by showing an event timeline for five complete threads of interface design discussion from the Drupal project. Though we only show five, the other threads analyzed exhibited similar patterns. The timeline shows the fluidity of topics (i.e. whether each message continues the topic of the previous message or changes it) along with visual indicators of when new alternatives were posted and when new participants joined the discussion. This figure offers interesting insights about the flow of conversation in the interface design discussions.

For example, within these threads, the proposal of design alternatives is distributed throughout the discussion, rather than batched at the very beginning reminiscent of

commonly applied face-to-face brainstorming techniques. This may be a consequence of the distributed nature of the discussion in time and space as well as its integration with an issue management system. It may also reflect the fact that participants are able to generate and submit new alternatives as a function of the ongoing discussion. However, an important consequence of having the alternatives distributed throughout the discussion is that some may be overlooked or even fade from community memory. As DU3 explained: “Even if 20 or 30 ideas get generated during the mailing list discussion, a few days on people will be discussing [only] one or two which might be the worst ones because they might be the most controversial.”

A related pattern is that the topic of the design discussion changes frequently. For example, in the first thread, there is a topic change after A3 (Alternative 3), A6, A8, and A11. Because of these topic changes there may be a reduced chance for these alternatives to be evaluated. In this case, A8 and A11 didn’t receive any comments, and A3 only received one. Indeed, the lack of structure in the discussion sometimes prompts a participant to write a summary of the discussion to date, including the alternatives and opinions of those alternatives. One such summary is called out in the first thread in Figure 4. In this case, the participant wrote the summary mainly to compensate for the lack of awareness in the system. Twelve ideas were proposed and people were struggling with deciding which one works best. The summary reminded them of the goals and the description of each alternative.

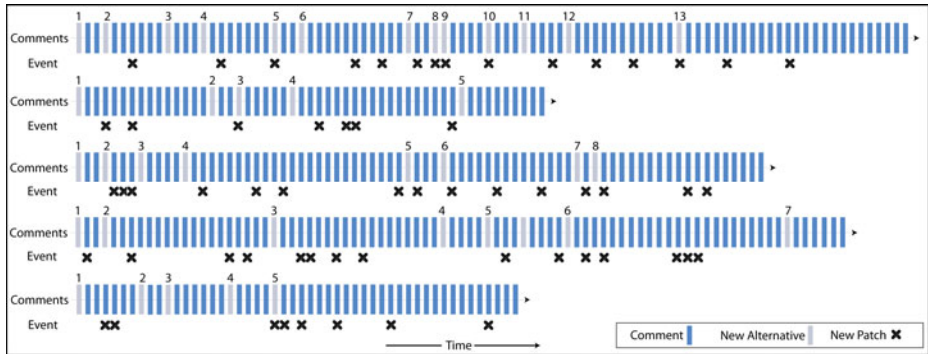


Fig. 5. The timeline from Figure 4 showing only the introduction of alternatives and patches

Another interesting pattern is that the majority of the participants joined during the first half of the discussions, but continued to join throughout. As the discussion grows, those who join later or otherwise do not keep up must review the messages to track the alternatives, the arguments for and against them, and the current consensus of the other participants. The common method for acquiring this information is to (re-)read the discussion to date. But, since this is time consuming, some participants will post irrelevant comments that hinder the flow of the discussion. As DD1 said: “one thing that gets very frustrating in this, it gets very frustrating when I’m involved in a long discussion and have been for the whole time, someone will often come in and just kind of jump in to the discussion and either drill it and say “Oh, this is such a great discussion I also noticed this other problem with this other issue or this other

thing” and people will go off on a tangent for two weeks talking about this other thing and we’ve gotten away from what the core issue is which is can be frustrating.”

To avoid losing bright ideas and to have a more organized discussion, the current systems should support better tracking of ideas. Highlighting the alternatives and connecting the messages that reference them can greatly aid participants. It could also reduce the time required to identify, compare, and consider the alternatives without having to sift through all of the textual comments in the discussion thread.

4.2 Integration of UI Design and Development Activities Is Essential

Designers and developers currently participate in the discussions through a centralized issue management system. This centralized venue helps designers in building trust and gaining merit by enabling them to interact with developers and exhibit their skills. Once they gain respect as a designer, they can more easily convince developers to implement their suggested improvements [5, 23, 27].

Integrating design and development activities also helps designers collaborate with developers. This collaboration is necessary for designers to receive feedback on the feasibility of their desired improvements [17] and for developers to be advised on the interaction design of their implementation. This iterative process of interface design and development is visible in current discussion threads. As shown in Figure 5 the proposed alternatives and submitted patches are distributed throughout the discussion threads, where each alternative is usually followed by a number of patches.

Any new interface for supporting interface design activities online should be fully integrated into the respective issue management systems. This will allow designers and developers to build mutual trust and collaborate more effectively.

4.3 Participants Need to Be Aware of Others’ Opinions Regarding Alternatives

Discussion participants typically demonstrate agreement by writing “+1” for a favored alternative or they simply state that they like the idea. In order to determine the current direction and the favored ideas, participants must read through the messages. Another option would be to ask others to clarify the current direction.

However, participants may have inconsistent perceptions about the direction of the issue. As DD1 said: “... *It’s often hard to figure out what is the current direction. That’s definitely hard to do. Often it takes getting someone to clarify it. And not everyone would clarify it the same way. If there are two people, and they’re each kind of pushing their own ideas, within a Drupal issue...and you were to go on IRC and ask each of them individually, “So what’s the current direction?”, you’ll get two very different answers.”*

The current issue management systems lack a formal way of expressing one’s preferred idea and visualizing others’ preferences. The absence of a mechanism to share opinions can hinder the consensus building process. Today, the consensus building process can be lengthy and it can be difficult to determine whether consensus has been reached at all. As DU3 said: “*People can keep on arguing the point, long after the decision was made [...] The nature of the way that many online discussions work is that they let the discussion continue [indefinitely]. That’s the main difficulty.”*

Implementing a mechanism to share preferences and formalize the consensus building technique (e.g. a voting system) may help facilitate the decision-making process. Also, it will be effective to highlight the alternative that has the consensus so far. Bringing the consensus to light can help developers determine which alternatives need to be implemented to improve the project or further inform the discussion.

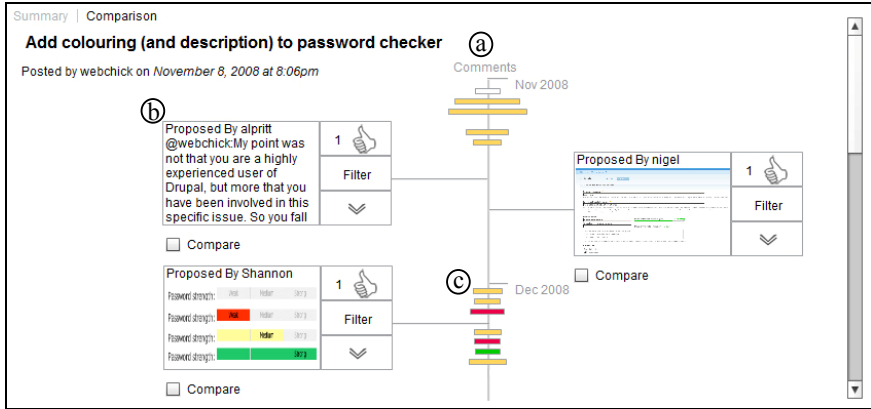


Fig. 6. The main screen consists of an interactive visual timeline, highlighting alternatives and offering an abstract summary of the comments. (a) The timeline shows the chronological order of comments and alternatives. (b) The alternatives are shown in callouts so designers can easily track them. (c) The comments are represented by a rectangle whose width corresponds to the length of the comment. The rectangles are colored based upon their affective tone.

5 IdeaTracker

We describe how we translated our implications into the implementation of an interactive visualization tool for reviewing online interface design discussions - IdeaTracker. Our tool can be used by designers and developers participating in a design discussion as well as facilitators who may join a discussion to facilitate the decision-making process. Our system was developed through an iterative design process, starting with four different prototypes that addressed the challenges identified in our study. An informal user study was conducted on these prototypes. For the study, each prototype was seeded with data from an actual design discussion. Four users representative of our target audience were recruited and asked to perform similar tasks (e.g., identify the idea that reflects community consensus) with each prototype and the existing interactive forum interface. The users were then asked to explain the strengths and weaknesses of each prototype. From the results, we implemented our prototype of IdeaTracker. We first discuss the main interface components of our system and then illustrate its value through a user scenario. All of the figures illustrating the use of our system are based on data imported from an actual design discussion

in Drupal where participants are proposing and debating alternatives for a revised password checker. To facilitate use and learning of the interface, all interactive controls in IdeaTracker have a tooltip which explains their functionality. For the visual elements, the user can access a short description of each element via a context menu.

5.1 Tracking Alternatives

IdeaTracker's main screen consists of an interactive visual timeline, highlighting the alternatives proposed in the current thread of discussion and offering an abstract summary of the posted comments (Figure 6). The timeline illustrates when a specific comment or alternative has been posted (Figure 6a). This timeline allows participants to gauge the amount of activity that has occurred within a specific timeframe as well as the overall progress and pace of the discussion. The alternatives are shown in separate callouts so designers can easily identify and track them (Figure 6b). If an alternative has an attached screenshot, the screenshot is shown in the callout; otherwise, the first few sentences describing the alternative are shown. All other comments are represented by a thin rectangle, with the width of the rectangle corresponding to the length of the comment (Figure 6c). The comments are colored based upon their affective tone. If a comment has a negative tone, the rectangle is colored red. If it has a positive tone, the rectangle is colored green. If the comment has both positive and negative words, then it is colored yellow. The number of negative and positive words is computed by looking up each word in a commonly used dictionary. This color coding allows designers to quickly assess the community opinion of a certain alternative. Designers can easily skim through the comments related to a specific alternative without having to read the text of each message. To aid in exploring alternatives and the comments regarding each alternative, three interaction mechanisms have been implemented:

Expand Alternatives and Comments: The user can select the expand/collapse button next to each alternative and read the entire post explaining the proposed alternative (Figure 7). Also, hovering over each comment representation will open a window containing the first few sentences of that comment (Figure 8).

Filter Unrelated Comments: To examine the comments related to a particular alternative, designers can press the filter button next to the alternative. This dims all of the representations that do not reference this alternative (Figure 9). This interaction isolates the pros and cons of an alternative pointed out by other designers. It also aids in detecting the alternatives that have received insufficient or controversial discussion.

Link to the Original Post: The user may want to read the original post corresponding to a comment or an alternative. To make this interaction possible, a link is provided in both expanded versions which redirect the user to the original post corresponding to that particular alternative or comment. Also, the title of the issue at the top of the main screen links to the original discussion thread.

Fig. 7. The user can press the expand button to reveal the entire post explaining the proposed alternative

Fig. 8. Placing the cursor over a comment's representation will open a window containing its first few sentences

Fig. 9. Users select 'filter' to dim comments unrelated to the alternative. Here selecting filter for the alternative in the top left shows one message referencing it in context of the discussion.

5.2 Comparing Alternatives

To compare different ideas, IdeaTracker offers a comparison view. Users select the ideas they would like to compare by clicking on the check-box at the bottom of each idea. Then, selecting the compare link will redirect them to the compare view (Figure 10). This view shows a timeline for each idea and the representations of comments referencing those ideas are available on the timelines for comparison. To provide users with a reference point for comparison, all of the comments are shown under each idea. But, the comments that are not related to a particular idea are dimmed.

5.3 Voting System

A voting system has been implemented in IdeaTracker to aid designers in promoting and reaching consensus. The number of votes for each alternative is shown on the vote button next to the alternative. Hovering over the vote button will show the list of people who voted for the idea. The user can vote for an alternative by clicking on the vote button. If the user clicks the vote button, the number of votes for that idea will increase by one and the vote button will be highlighted to indicate which idea the user

has voted for. To synchronize IdeaTracker with the original issue, a comment will be automatically posted to the original issue on behalf of the user stating that the user favors that particular alternative. Conversely, if a user posts a comment using the common notion of “+1” for an alternative in the original discussion thread, the number of votes for that alternative will be updated in the IdeaTracker.

Through IdeaTracker each user can only vote for one idea. If a user votes for a different idea, her initial vote will be re-assigned to the new idea. This feature enables users to retract their votes if a better idea is proposed or an existing idea is refined. This way IdeaTracker reflects the participants’ recent views about the proposed alternatives. To promote awareness of the current consensus, the idea with the highest number of votes is highlighted in the system (Figure 11).

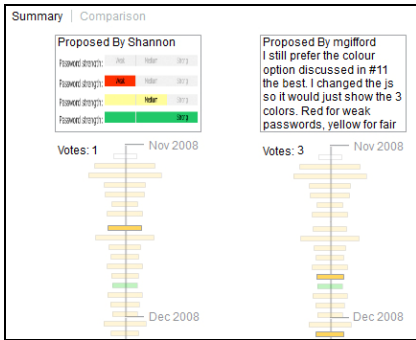


Fig. 10. The compare view shows a timeline for each selected alternative and the comments referencing them. All of the comments are shown under each alternative, but the comments unrelated to the particular alternative are dimmed.

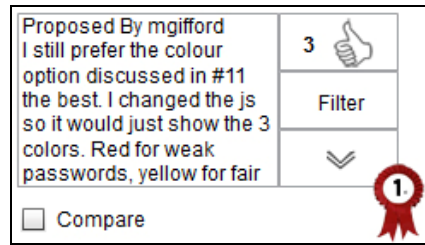


Fig. 11. To promote awareness, the alternative currently with the most votes is highlighted

5.4 User Scenario

Bob is a UI designer who contributes to open source projects in his spare time. Looking through the usability issues in Drupal, he finds an issue about improving the usability of Drupal’s password checker. He launches IdeaTracker and enters the URL corresponding to the discussion thread for that design issue.

He first wants gain awareness of the proposed alternatives and their pros and cons as pointed out by other participants. He quickly scans the list of alternatives highlighted on the main screen and notices the idea posted by Lisa that suggests borrowing the design of Google’s password checker. Bob then expands the idea to read it in detail and selects the filter button to identify the comments that reference that idea. He immediately notices a negative comment colored in red and hovers the cursor over the comment to read it. The comment has been posted by Mark who thinks that copying and pasting a design from Google will spoil Drupal’s trade mark. Bob agrees, so he continues to scan the alternatives to determine if someone else has proposed a better

solution. At the end of the discussion, he finds Anne's idea. Anne's idea is a tweak of Lisa's idea. She suggests that instead of using red, orange, and green to indicate a weak, medium, and strong password, they can use different shades of green. Bob decides to compare Anne's and Lisa's ideas. He selects the compare checkbox next to these ideas, selects the compare link, and is redirected to the compare view. He reads the comments posted regarding each idea and compares them. He decides that he likes Lisa's idea more than Anne's. He returns to the main screen and votes for Lisa's idea. A comment is generated on Bob's behalf and posted to the original discussion thread indicating that Bob is in favor of Lisa's idea.

In contrast, using only the current interface, Bob needs to review each comment to identify the alternatives and their pros and cons. He then either needs to use another tool to create a summary of the discussion or rely on his memory. Comparing the two alternatives would also be challenging because it is difficult to isolate the comments that specifically address only the desired alternatives. Finally, since there is no running tally of "+1" votes, it is difficult to identify the currently favored idea.

6 Implementation

IdeaTracker is fully implemented and its interface was written in ActionScript using Adobe Flex 3 interface framework. The software consists of two layers: the data and visualization layers. The data layer parses the collected data and translates it into an internal format understood by the visualization layer. The data layer receives the data in XML format. The XML data consists of a set of <comment> tags, and each <comment> tag should have <author>, <content>, <date>, and <image> tags.

When the user launches IdeaTracker to view a particular discussion thread, an adaptor component parses the html source of the thread and converts it to the XML format readable by the data layer. The data layer is independent of the html format and only depends on the XML format. In order to apply IdeaTracker to design discussions on other interactive forums, an adaptor component needs to be written that translates the html source of that forum to the XML format readable by our tool.

The data layer processes the XML file to find all the posts. In order to find the alternatives we use two heuristics: (i) we consider posts with image attachments as alternatives and (ii) we consider posts that have been referred to by other posts as alternatives. Our testing indicates that these heuristics accurately detect most alternatives in a thread of discussion. In future work, discussion participants could be allowed to insert a simple tag in their comment stating they are posting an alternative.

After detecting the alternatives, we use natural language processing techniques to infer the comments related to each alternative and the tone of each comment. To find the comments related to an alternative, we look for certain key phrases participants commonly use to reference a comment, for instance: "#34" to refer to comment number 34 or "@Lisa" to refer to the latest comment posted by Lisa.

In order to determine the affective tone of a comment, we find the number of negative and positive words in the comment using standard online dictionaries. Based on the percent of negative and positive words, we assign respective values to that comment, which is used by the visualization layer for color coding.

7 Preliminary Evaluation

We conducted a qualitative evaluation to assess design choices and gauge initial user reactions to IdeaTracker. The evaluation was performed using the implementation of our tool as described in the previous section. It involved eight designers and six developers who actively contribute to interface discussions in Drupal and Ubuntu. The evaluation started with an introduction to IdeaTracker and a demonstration of its main features. Afterward we asked participants about their perceptions of the overall direction and different features of the tool (e.g. what do you think about showing the ideas in separate callouts, providing an abstract visualization of the comments, and using color codes for the affective tone of the comments?) and encouraged them to respond openly. Each session lasted about thirty minutes.

Overall, the participants reacted positively to IdeaTracker. All of the participants appreciated the visual separation of ideas from the other comments and being able to filter comments related to a particular idea. For example, one Drupal designer said: *“The most useful feature to me is the callout of the major ideas that cuts through all lot of the crufty comments”* while another said *“I like this sort of compressing it... here is some big comments, here is a bunch of small comments, and if they are generally in favor or not, sort of at a glance as an overview is very cool.”* Most of the participants appreciated having access to an abstract visualization of the comments and felt we were using reasonable decision rules for identifying ideas and filtering comments. Participants also appreciated the fact that IdeaTracker was seeking to complement the existing issue management systems rather than trying to replace them.

The evaluation also highlighted several opportunities for improvement. For example, some of the participants were unsure of the utility of the idea-centric comparison view. Instead they preferred the ability to filter comments based on user id, thereby allowing them to see the comments that one user made across all of the ideas. Participants also expressed that the content of the negative and positive arguments for an idea was more important than the number of votes. For example, as one Drupal designer said: *“Often there are issues though where an idea has lots and lots of “likes” until one person discovers why it shouldn’t be done...”* It may therefore be useful to extract the arguments for and against an idea and represent them within the main visualization. Most of the designers were concerned about accuracy of coloring comments based on affective tone and suggested to color comments based on their type (e.g. code review or patch). Participants also asked for more information to be included in the visualization of each comment (e.g. who posted the comment).

8 Conclusions and Future Work

With the emergence of open source software and the geographic distribution of many design teams, more interface design discussions are occurring online. The discussions are carried out using typical interactive Web-based forums, which lack support for the unique nature of design discussions. This paper makes two contributions. First, we conducted a study examining the challenges faced by participants when using the current interactive forums. The study included analyzing the discussion content from two popular open source projects and conducting interviews with active participants

in these projects. From the results, we identified key challenges of using these types of systems and implications for how they could be addressed. Second, we built a new interactive visualization tool called IdeaTracker to demonstrate how to manage online design discussion more effectively. The interface offers specific support for tracking alternatives, comparing alternatives, and building consensus, all of which were identified as significant challenges in our study but which are not supported by existing interactive forum models. The interface was also implemented such that it can collect actual data from existing discussion threads for several open source projects and can be adapted for others.

We see at least three promising directions for future work. First, we want to conduct a longitudinal study to compare the impact of our system on the process and outcomes of design discussions relative to the use of existing interfaces. Second, we would like to test different techniques for building consensus, for example, voting for versus ranking the ideas. Finally, we would like to integrate more sophisticated machine learning techniques to better identify key elements of the design discussions such as identifying the alternatives and the affective tone of a discussion.

References

1. Axure, <http://www.axure.com>
2. Balsamiq, <http://www.balsamiq.com>
3. Pencil, <http://www.evolus.vn/pencil>
4. Arias, E., et al.: Transcending the individual human mind - creating shared understanding through collaborative design. *ACM TOCHI* 7(1), 84–113
5. Bach, P.M., DeLine, R., Carroll, J.M.: Designers wanted: participation and the user experience in open source software development. In: *Proc. ACM Conference on Human Factors in Computing Systems*, pp. 985–994 (2009)
6. Baddeley, A.D.: Working memory. *Science* 255, 556–559 (1992)
7. Barcellini, F., et al.: A Study of Online Discussions in an Open-Source Software Community. In: Besselaar, P., et al. (eds.) *Communities and Technologies 2005*, pp. 301–320. Springer, Netherlands (2005)
8. Benson, C., Muller-Prove, M., Mzourek, J.: Professional usability in open source projects: GNOME, OpenOffice.org, NetBeans. In: *CHI 2004 Extended Abstracts on Human Factors in Computing Systems*, pp. 1083–1084 (2004)
9. Brown, D., Chandrasekaran, B.: *Design problemsolving: Knowledge structures and control strategies*. Morgan Kaufmann, San Francisco (1989)
10. Deshpande, A., Riehle, D.: The Total Growth of Open Source. In: Russo, B., et al. (eds.) *Open Source Development, Communities and Quality*, pp. 197–209. Springer, Boston (2008)
11. Everitt, K.M., et al.: Two worlds apart: bridging the gap between physical and virtual media for distributed design collaboration. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 553–560 (2003)
12. Faridani, S., et al.: Opinion space: a scalable tool for browsing online comments. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 1175–1184 (2010)
13. Frishberg, N., et al.: Getting to know you: open source development meets usability. In: *CHI 2002 Extended Abstracts on Human Factors in Computing Systems*, pp. 932–933 (2002)

14. Janssen, J., et al.: Online visualization of agreement and discussion during computer-supported collaborative learning. In: Proc. Conference on Computer Supported Collaborative Learning, pp. 314–316 (2007)
15. Liapis, A.: Synergy: a prototype collaborative environment to support the conceptual stages of the design process. In: Proc. International Conference on Digital Interactive Media in Entertainment and Arts, pp. 149–156 (2008)
16. Muller-Prove, M.: Community experience at OpenOffice.org. *Interactions* 14(6), 47–48 (2007)
17. Muller, M.J.: Retrospective on a year of participatory design using the PICTIVE technique. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, pp. 455–462 (1992)
18. Myers, B.A., Rosson, M.B.: Survey on user interface programming. In: Proc. ACM Conference on Human Factors in Computing Systems, pp. 195–202 (1992)
19. Nicholas, D.M., Twidale, M.B.: The usability of open source software. *First Monday* 8(1) (2003)
20. Nicholas, D.M., Twidale, M.B.: Usability processes in open source projects. *Software Process: Improvement and Practice* 11(2) (2006)
21. Olson, G.M., et al.: Small group design meetings: an analysis of collaboration. *Hum.-Comput. Interact.* 7(4), 347–374 (1992)
22. Sack, W.: Conversation map: a content-based Usenet newsgroup browser. In: Proc. ACM Conference on Intelligent User Interfaces, pp. 233–240 (2000)
23. Terry, M., Kay, M., Lafreniere, B.: Perceptions and practices of usability in the free/open source software (FoSS) community. In: Proc. ACM Conference on Human Factors in Computing Systems, pp. 999–1008 (2010)
24. Thomas, M.P.: Why Free Software has poor usability, and how to improve it (2008), <http://mpt.net.nz/archive/2008/08/01/free-software-usability>
25. Twidale, M.B., Nichols, D.M.: Exploring Usability Discussions in Open Source Development. In: Proceedings Hawaii International Conference on System Sciences, January 3-6, vol. 07, p. 198. IEEE Computer Society, Los Alamitos (2005)
26. Vora, P., Helander, M.: A review of design models and a proposal for a taxonomy of design. In: Helander, M., Nagamachi, M. (eds.) *Design for Manufacturability: A Systems Approach to Concurrent Engineering and Ergonomics*, pp. 78–90 (1992)
27. Zilouchian Moghaddam, R., Twidale, M., Bongen, K.: Open Source Interface Politics: Identity, acceptance, trust, and lobbying. In: CHI 2011 Extended Abstracts on Human Factors in Computing Systems (2011)

What You See Is What You (Can) Get? Designing for Process Transparency in Financial Advisory Encounters

Philipp Nussbaumer and Inu Matter

University of Zurich, Department of Informatics, Binzmuehlestrasse 14,
8050 Zurich, Switzerland
{nussbaumer,matter}@ifi.uzh.ch

Abstract. In this paper, we report on a study to establish process transparency in service encounters of financial advisors and their clients. To support their interaction, we implemented a cooperative software system for tabletops, building on transparency patterns suggested by the literature. In evaluations, however, we found that our design did not improve the perceived transparency and comprehensibility. Introducing the IT artifact into advisory failed to enhance the client’s overall experience and even seemed to negatively influence the client’s perception of the advisory process. Using the representational guidance of depicting the process and its activities as a navigable, interactive map made clients believe that interactions with their advisor were restricted to the system’s functionality, thus expecting that what they see is all they *can* get.

Keywords: process transparency, collaboration, advisory, tabletops.

1 Introduction

In the changing market environments of the last several years, an increasing number of financial service providers (FSPs) has turned to individualized financial advisory services as a competitive differentiator. For Swiss banks, however, [1] found that implementation and application of such services are still in their infancy and that customers are inherently dissatisfied with their FSP’s service provision. A main point of criticism thereby is related to a lack of transparency and comprehensibility of advisory services, i.e., the activities performed therein, their interrelations as well as their results. In service encounters, the client often perceives her advisor as a “black box”, collecting client information as an input and returning investment recommendations as an output, usually without revealing a consistent line of reasoning. Thus, it is difficult for clients to relate their voiced needs (i.e., their problem space) to the advisor’s recommended strategy and products (i.e., the solution space) [2]. Additionally, the information asymmetry between the actors (with the advisor typically being more knowledgeable) and the potential interest asymmetry (advisors exploiting information asymmetry to take advantage on their clients) lead to further distrust and thereby impact client satisfaction.

In this paper, we argue that establishing process transparency in service encounters may alleviate these issues, i.e., allow for client understanding and comprehensibility

while decreasing information and interest asymmetries. The notion of process transparency has been discussed in different research fields with various meanings and levels of concreteness. A suitable conceptualization can be found in organizational literature, where process transparency is often related to internal and external process communication. As such, transparency has been argued of being related to customer satisfaction, as the transparent communication of processes enables clients to appreciate the company's activities and efforts as well as to identify their role in service provision. On a more specific level, in CSCW research, process transparency corresponds to communication processes in work groups and has been shown to influence the organization of communication and cooperation processes (e.g., [11], [12], [13]).

To support process transparency in service encounters based on these notions, we designed a collaborative tabletop artifact providing visualization and simulation functionality for the main activities of investment advisory, prominently featuring a navigable depiction of the process and its activities (e.g., needs elicitation, risk analysis, definition of investment strategy). The design's rationale was to present and allow control of the organization's advisory process and its activities, thereby mediating advisor-client communication and furthering the client's understanding.

To measure the effects of increased process transparency on the interaction of client and advisor as well as the client's understanding and satisfaction, we compared the traditional setting (advisor using pen & paper) with the IT-supported setting (advisor and client cooperatively using the tabletop artifact) in controlled within-subject evaluations. Surprisingly, perceived transparency and comprehensibility as well as satisfaction were rated the same or even lower for the IT-supported setting compared to the traditional setting.

We found several explanations for the design failing to improve the client's overall experience. Most prominently, we found that clients felt rather restricted by the system, taking the depiction of the process and its activities as the boundaries of advisory along the lines of "if you cannot see it, you cannot get it". Despite of having based our design considerations on comprehensive explorative research and having evaluated the resulting designs with domain experts, we underestimated the side effects of process transparency on advisor-client interaction.

2 Transparency Issues in Financial Advisory

Today, the most promising strategy of financial service providers to differentiate against competitors is to offer highly personalized services. These cannot easily be compared or imitated due to their dynamics and complexity [3]. However, since the fundamentals of such services have not yet been established, FSPs have been counteracting cost pressure by optimizing their advisory services towards efficient and effective product sale rather than individualized advisory. As a consequence, the quality of advisory services has been perceived as rather dissatisfying or even inappropriate for customers (e.g., [4], [1], [5], [6]).

In an exemplary investment advisory consultation of a Swiss bank, the client and the advisor meet in a designated consultation room. In the case of prospect clients, they meet for the first time, so the advisor has minimal information about the specific needs of his *vis-à-vis*. Thus, for the first few minutes he will engage in small talk to

gather basic information about the client (financial situation, needs and wishes), taking notes on his notepad. The advisor then typically presents the bank's generic advisory process, which he will use to optimize the client's financial situation and to help her in achieving her goals. Throughout the remainder of the encounter, the advisor tries to gather as much information about the client's financial situation, her risk preferences, investment experiences as well as her interests in particular asset classes. Building upon this information, he will then suggest an investment strategy that proportionally attributes the client's investment to different asset classes (e.g., shares, bonds, money market). After some iterations of adapting this strategy to the client's preferences (e.g., increasing the amount of bonds and decreasing the amount of shares), the first encounter is finished (typically after up to 90 minutes). The advisor will propose to prepare a product portfolio for the agreed strategy, which will be either sent to the client (including material for establishing the contract) or discussed in a subsequent encounter.

Regarding such an advisor-client interaction, several characteristics can be found that are detrimental to perceived advisory quality. Most prominently, such encounters are inherently impacted by information asymmetry and interest asymmetry, problems that are well established in scientific literature in context of the 'principal-agent problem' [7, 8]. Information asymmetry results from the customer being generally less knowledgeable than the advisor – thus, she cannot be sure whether the advisor actually gathers and provides all relevant information and recommends appropriate solutions for her financial needs. The relation between customer and advisor can be additionally strained by conflicts of interests. Advisors might exploit information asymmetry by, e.g., superficial information gathering and provision or, even worse, recommending products that are unsuitable for the specific customer's needs but profitable in terms of fees.

In a comprehensive study of advisory practice in Swiss banks, [1] found that clients are quite aware of these problems. As a result, they do not consider financial advisors as being very trustworthy – also, they perceive the advisor's knowledge about market trends or even the bank's products to be rather limited. All in all, clients are not very confident that financial advisors present adequate solutions to their needs. Many of the problems found in advisory encounters relate to a perceived lack of transparency, i.e., the absence of comprehensible and coherent advisory schemes that allow verification of the progression and results. Not recognizing the underlying rationales of advisory and its activities, clients also perceive advisory encounters as lacking personalization [1].

3 Design of the Prototype System

Analyzing the status quo of financial advisory from interviews and discussions with advisors and clients (including interviews of 21 advisors from 19 FSP, client focus groups totaling 28 participants as well as a client survey with 136 participants; see [1]), we designed problem scenarios of typical financial advisory encounters to derive generic design considerations for comprehensible and interactive service encounters between advisor and client [9, 10], including the notions of cooperation, process transparency, information transparency as well as cost transparency.

In the following, we will focus on the concept of process transparency. Building on the issues identified in our field work, we define process transparency as the degree of the client being able to follow and comprehend the performed activities (*what* constitutes an activity and *why* is it performed) and their succession in advisory. We will discuss the according requirements, the designs we chose to satisfy them as well as their implementation for a multi-touch tabletop device.

3.1 Design Requirements

From discussions with advisors and clients and observations of work practice, we derived the following basic requirements to enhance transparency of advisory encounters:

DR.1 Process awareness: Clients commented about the advisor being a “black box”, i.e., interaction between advisor and client being somewhat unpredictable regarding the information gathered by the advisor and how they influence the output of advisory. We therefore suggest making the process of advisory *visible*, i.e., increasing the client’s understanding by transparently presenting the performed activities and associated actions. This also involves structuring the succession of activities and specifying defaults for their execution.

DR.2 Process adaptability: In discussions with advisors and managers, we found that advisors do not favor the rather rigid process guidelines, which organizations try to establish in their efforts to standardize advisory activities. Processes and activities visibly mediating advisor-client interaction therefore have to be adaptable, i.e., they must feature multiple starting points and offer the possibility to change their order; thereby, both the advisor and the client should be enabled to keep track of the progress of the advisory process.

DR.3 Shared information space: Addressing the information and interest asymmetry between advisor and client requires the provision of transparent information access for both parties. This shared information space should support communication and interaction of advisor and client while also allowing for process awareness and adaptability as discussed above.

3.2 Design Rationales

Research on Computer Supported Cooperative Work (CSCW) and Computer Supported Collaborative Learning (CSCL) has been engaged in applying mechanisms of process transparency to support and enhance organization of communication processes. Almost twenty years ago, [11] already stated that “computer support of cooperative work should aim at supporting self-organization of cooperative ensembles as opposed to disrupting cooperative work by computerizing formal procedures” (p. 17). They acknowledged, however, that the organizational models should be made accessible to the users by the system, supporting the user in interpreting procedures and evaluating their rationales and implications.

For group work, [12] argues that the underlying communication and cooperation processes are not always clearly defined and comprehensible and thereby lack transparency regarding the status of the group work relative to the overall process; in such

cases, failing to establish process transparency may have undesirable outcomes on communication [13]. For similar reasons, research also suggests that members of computer-supported work groups have to explicitly agree on communication and cooperation processes [14].

External representations have been studied in the context of learning and problem solving tasks, showing that the nature of representation may influence the conception of the problem and hence ease the finding of an appropriate solution (e.g., [15], [16]). For collaborative processes in learning, [17] investigated the influence of representational guidance by comparing different tools for constructing representations of evidential models on collaborative learning processes and outcomes. They argue that external representations play at least three roles in situations of groups using shared representations in constructive activities (p. 473):

- (1) Initiating negotiations of meaning: when constructing or manipulating shared representations and trying to obtain agreement, members of a group have to explicate and negotiate the representations' meanings and their shared beliefs.
- (2) Being a representational proxy for deixis: collaboratively constructed representations may serve as an easy way for participants to refer to previous ideas and facilitate subsequent negotiations, thereby increasing the conceptual complexity that may be handled in a group's interaction.
- (3) Enabling implicitly shared awareness: shared representations may also serve as an external memory of the collaborating group, reminding the group members of previous ideas and comments.

Testing different representations with pairs of participants, [17] found that representations have impacts on learners' interactions and may differ in their influence on subsequent collaborative use of the knowledge being manipulated; thereby, visually structured representations can provide guidance for collaborative learning that is not afforded by plain text. They also speculate that graph representations will be most useful for gathering and relating information.

In [18] two approaches of establishing representational guidance are discussed, varying in their degree of process structure. *Maps* strive to only provide a basic means of orientation while not constricting the actual enactment of a process. In [19] such an approach was used in the context of process knowledge learning; they provided their students with a graphical hypermedia-based process representation to support cooperative process enactment; the representation contained associated materials with which the users could interact while systematically carrying out the process they were learning. In [20], Carrell et al. found that the usage of graphical process models during the preparation of collaboration leads to more knowledge exchange and integration and stronger individual and collaborative use of the software platform.

Kienle [18] uses diagrammatic representations for a software system supporting collaborative learning processes. She thereby suggests that navigable models are apt to give orientation and structure in the sense of representational guidance; as the model is always present for the user, it also should be internalized more easily.

In contrast to maps, which have the goal to give only basic orientation of communication and cooperation processes, (cooperation) *scripts* consist of detailed instruction sets of how a group should interact and collaborate to solve a given task. Being highly structured, such processes (and their representation in software systems) are

rather inflexible and run the risk of not fitting the processes already established in groups, which in turn might lead to non-acceptance [21].

Similar to [18], in our design we follow the map concept and address process awareness (DR.1) by depicting the advisory process as a fixed-positioned, navigable diagram, which is always present and can be controlled by both the advisor and the client (see Figure 1). As such, we also built our design on the different roles of external representations discussed by [17], enabling the advisor and client to discuss and reflect upon the process, allowing them to select and revisit activities at any time and track progress (DR.2), as well as helping to establish a shared information and activity space (DR.3) that may serve as an external memory. To enhance the system’s affordance for the users, we based the information design (e.g., pie charts, risk-return-graph; see Figure 1) on actual information material used for investment advisory in Swiss banks.

In contrast to the representations discussed by [12, 18] and [17], however, the design does not allow manipulating or changing the representation of the process itself. This is due to mainly two reasons: advisory processes typically consist of three



Fig. 1. Basic design of the system’s front end

generic process steps, i.e., (1) capturing and discussing the current situation, (2) finding and configuring optimizations, (3) implementing the optimizations; these steps are related to specific activities, which are generally accomplished in a specific order (e.g., for the optimization phase: discussing the client's risk profile and constructing a suitable strategy). Also, most of the activities found in financial advisory are interdependent (e.g., the strategy being a function of the client's risk profile and her preferences, which are discussed in the first and second process step). Thus, configuring these specific process steps and activities regarding their order might not be meaningful in most cases.

Secondly, the advisory processes are subject to compliance (standardization) as well as legal regulations (due diligence), constraining an advisor's freedom to "skip" activities. From the client's point of view, we also wanted to depict the "standard" process and its associated activities, as to give her an overview of the different advisory steps and associated activities. We did, however, consider it to be important that the advisor and client could themselves decide on the order of activities, allowing them to use the system's functionalities as needed. Therefore, the design allows ignoring interdependencies of activities, e.g., permitting to create an investment strategy without having configured the client's risk profile.

- (1) **Navigable process map**, indicating the current process step and activity as well as the overall progress: the highlighting indicates that the current activity is associated to finding an appropriate investment strategy (asset allocation) based on the information gathered in the previous steps.
- (2) **Shared information/activity space**: in the current activity (definition of investment strategy), the shared information space allows the advisor and client to collaboratively specify and adapt the client's asset allocation (interactive pie chart on the bottom right) while simultaneously illustrating the impacts on risk and return (graph on the bottom left).
- (3) **Advisor "Cockpit"**: shortcuts to navigate the process and its activities.
- (4) **Client "Cockpit"**: allows the client to display additional information for the specific activity as well as to access the projected asset growth based on the entered information (as depicted in Figure 2).

3.3 The SurFinance Prototype

To implement our prototype, we decided to use a multi-touch tabletop device (Microsoft Surface), so the advisor and client could simultaneously interact with the shared information and activity space without explicit handovers. Also, equipping a situation previously being equipped with pen & paper only, we assumed that a tabletop would be perceived as less intrusive and less disruptive for social interactions. In the newly designed encounter, the actors seat themselves at the tabletop device (see Figure 2 for an overview), which supports them in accomplishing the most important and complex activities (needs elicitation, risk profiling, strategy development, product selection).

While engaging in initial small talk, the advisor is enabled to transparently add the client's needs into an area at the center of the screen, assuring the client that her wishes and needs are taken seriously. To stimulate the client in thinking of additional needs and wishes, pictograms of basic categories (planned purchases, education, and housing) are readily available. Wishes and needs may be detailed with costs and

contextualized with a timeline to express the desired period of goal fulfillment. Using the client's financial information, a projection of the potential growth of wealth is added to the timeline, allowing for an assessment whether the client's goals may be accomplished. In such discussions, the advisor acts as a coach, who strives to enable the fulfillment of the client's needs by mapping them to appropriate financial strategies and products.

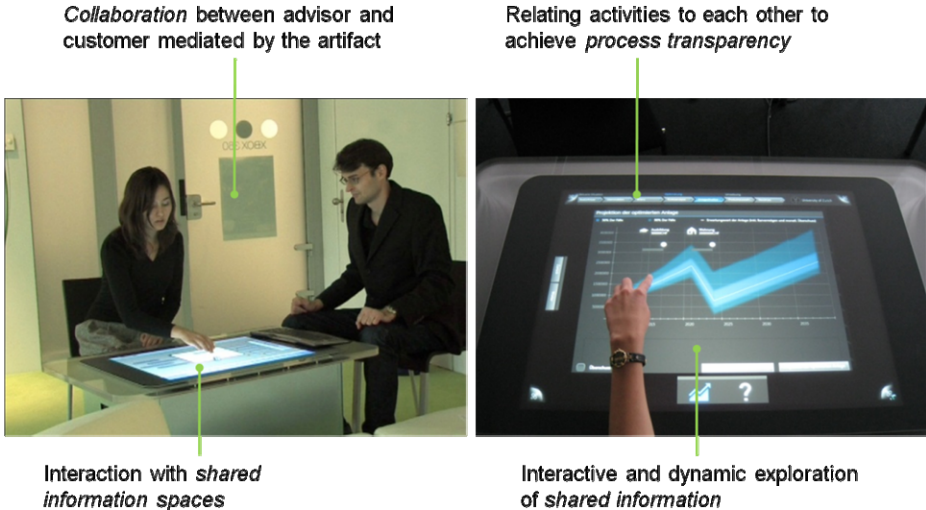


Fig. 2. SurFinance prototype

Collaboratively using the artifact, client and advisor are enabled to jointly define investment strategies that transparently include the defined needs and goals. The dynamic visualization enables the advisor to comprehensibly argue for or against specific strategies, while the client can immediately track the impacts on her financial situation. As an overview of all performed activities is provided at any given time, the client may also refine and revise her data by directly navigating to the specific activity. Having agreed on a strategy, the client and advisor may directly implement it by selecting appropriate products, or the advisor – similar to the traditional setting – may prepare an appropriate portfolio for a follow-up encounter.

3.4 Hypotheses

Based on the previous discussion and our experience from exploratory research [1], we state the following hypotheses.

One of the main design goals of our prototype is to depict the advisory process and visualize its activities, allowing the clients to actively interact. We hypothesize that these novel possibilities increase the client's comprehensibility and understanding of advisory's contents and activities. We thereby implicitly assume that in traditional advisory situations – having to rely on the advisor's explanations and drawings – it is

difficult for clients to establish a detailed understanding. The according hypothesis reads as follows:

H1.1: *Using an IT artifact enabling process transparency, the comprehensibility of the order of activities as perceived by the clients is higher than that of the traditional financial advisory encounter.*

With our prototype, we aim at the client becoming an integrative co-producer of the advisory result by contributing to and interacting in activities, while immediately being able to assess their intermediate results. We therefore hypothesize that using the artifact increases the client's understanding of how and why the results came about:

H1.2: *Using an IT artifact enabling process transparency, the comprehensibility of the results as perceived by the clients is higher than that of the traditional financial advisory encounter.*

In focus groups and interviews, clients often argued that advisors tended to recommend their "standard" products without the clients being able to influence their decision making. We therefore hypothesize that the presence of an interactive, shared artifact should increase the perceived degree of being able to influence the process and its results compared to the traditional situation, which in turn should be perceived as more restrictive:

H2.1: *Using a shared IT artifact enabling process transparency, the degree of being able to influence the solution finding process as perceived by clients is higher than that of the traditional financial advisory encounter.*

In addition to a higher degree of influencing the process, we hypothesize that in the IT-supported setting the client is better enabled to actively participate in activities:

H2.2: *Using a shared IT artifact enabling process transparency, the client perceives a higher possibility to participate in activities than that of the traditional financial advisory encounter.*

As discussed above, surveys show clients being inherently dissatisfied with advisory services of their banks. Having identified transparency issues as a possible cause, we suggest that introducing IT-enabled process transparency should also positively affect the clients' satisfaction with the advisory encounter.

H3: *Using a shared IT artifact enabling process transparency, the client's overall satisfaction is higher than that of the traditional financial advisory encounter.*

4 Experimental Evaluation

We built our SurFinance prototype in two iterations. The first prototype was implemented by a group of four students as part of a Master's project, and already featured the basic transparency design reported in Section 3. To evaluate our design rationales, we discussed the prototype and its underlying concepts with representatives of four major Swiss banks. Their unanimously positive feedback encouraged us to enhance and functionally extend our prototype in order to evaluate it with real users. To get directions in revising the prototype and finding additional functional requirements, we conducted three focus groups of 15 domain experts in total (one focus group for

financial advisory experts, financial software developers and design experts each); based on their input, we built the second iteration of the prototype. In the following, we will report on the experimental evaluation of this prototype (as described in Section 3.3).

4.1 Experimental Design

The evaluation involved 12 clients, each of them performing an investment planning task in two different settings in a within-subject design. The participants were recruited by convenience sampling through postings on a university forum (each received 40 CHF), eight of them being university students of various study programs. The participants were between 21 and 50 years of age, with high proficiency in computer use (6 participants categorized themselves as being professional users, 5 as advanced users and only 1 participant reported to use IT only occasionally). Half of the participants were female, and five participants already were experienced with advisory.

Clients received a short introduction before the test sessions, including explanations about their time table, instructions about their task and financial profile. The planning task involved the investment of a specific amount of money from 250'000 to 480'000 CHF, while considering specific wishes and goals, e.g., purchasing an apartment. The clients were allowed to keep back their information until the advisor explicitly asked for it. To preserve the participants' privacy, they were provided with a profile that included key figures of their assumed financial situation. Each client participated in two test settings. One setting corresponded to the traditional (pen & paper) advisory that is typically provided in Swiss banks, the other setting involved the use of the SurFinance prototype. The participants were randomly assigned to either start with the traditional or the IT-supported test setting (50% of participants each). Test sessions of each setting were limited to 30mins.

The sessions were conducted by four financial advisors (three being male, one being female) of a Swiss bank. Their age was between 31 and 40 years. They had been practicing their job as financial advisors for four up to seven years. All of them classified themselves as being advanced IT users. Each of them received a 30mins hands-on training with the prototype system. In their briefing, the advisors were given two main instructions:

- (1) In the traditional test session (without IT), they were asked to perform advisory along their actual practice. Advisors were allowed to use all material they needed to advise a prospect client asking for support in investment planning. Prior to the evaluation, we therefore asked them to bring the according material with them.
- (2) In the test session featuring the SurFinance prototype, the advisors were required to use the artifact at least once in their session (providing interaction possibility for the client) but were free to decide at which point they would introduce the tool and in what order they would perform specific activities (need elicitation, risk profiling, etc.).

Each advisor performed three traditional advisory sessions (without IT) as well as three IT-supported sessions using the SurFinance prototype.

4.2 Data Collection

The tests were conducted on two days at the end of July and beginning of August 2010. After their trials, clients received a quantitative questionnaire and were debriefed in semi-structured interviews (having an average duration of 30mins); advisors were asked to provide qualitative feedback in semi-structured interviews (average duration of 60mins). The semi-structured interviews covered the following main aspects: disturbing moments, comprehensibility of the advisory sessions, collaborative activities and interactive moments, advantages and disadvantages, preferences regarding the advisory setting (traditional vs. IT-supported) and additional comments.

Our quantitative questionnaire included items to test our hypotheses as well as demographic items (age, gender, education, advisory experience, IT skills). As we are not aware of any standardized items for comprehensibility (H1.1 and H1.2) of (advisory) processes and the succession of activities, we measured comprehensibility with the following two items:

- “I could understand at any time why the activities of the advisory session were following a specific order.” (Comp1)
- “I do understand how the results of the advisory session have been achieved.” (Comp2)

To investigate the client’s perceived influence on the advisory process (H2.1) and on her ability to perform actions (H2.2), we used the following two items:

- “Overall, I was able to influence the solution finding process of the advisory session.” (IoAP)
- “Overall, the advisory situation enabled me to participate in activities.” (PA)

For hypothesis H3 we used items of the Yield Shift Theory of satisfaction (Briggs et al. 2008).

Each item was measured once for each advisory session (traditional and IT-supported advisory) with a seven-point Likert-scale (from 1 = “I strongly disagree” to 7 = “I strongly agree”).

4.3 Results

The data from our evaluation were tested with two-sided t-test for paired samples with differing variances.

The perceived comprehensibility of the order of activities (Figure 3a; avg. traditional setting (TS) = 5,08; avg. IT-supported setting (ITSS) = 5,00) as well as the comprehensibility of the results (Figure 3b; avg. TS = 5,42; avg. ITSS = 4,58) were rated lower for the IT-supported setting than for the traditional setting. The t-test, however, did not show any significant difference between the two settings. We can therefore neither support nor falsify hypotheses H1.1 and H1.2.

At the debriefing interviews, clients brought forward several reasons for their ratings on comprehensibility. Two participants found that the advisor’s explanations regarding the process and its activities were better in the traditional setting than those received in the IT-supported setting. They reported that it was unclear to them how the charts and results of the IT artifact came about. Advisors also reported to having

had difficulties in explaining visualized information, especially when charts contained multiple information dimensions.

One client and one advisor perceived the conversation’s pace and progress of the IT-supported setting as too fast. Another three clients experienced “information overflow”; while one was overwhelmed by the amount of information channels (IT-artifact *and* advisor), another found that the IT artifact required too much knowledge. Advisors pointed out that this resulted in clients asking more specific questions, which were difficult to answer because of the clients’ lack of knowledge.

In the traditional setting, two clients reported to be overwhelmed by information (e.g., investment possibilities) provided by the advisor.

Five clients argued that in the traditional setting the conversation with their advisors was more consistent and “smooth” as opposed to the IT-supported setting, where conversation was “interrupted” by the use of the IT artifact.

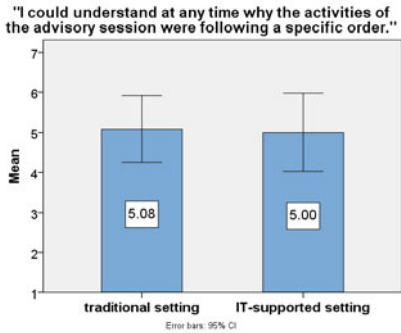


Fig. 3a. Mean Comp1 (error bars: 95% CI)

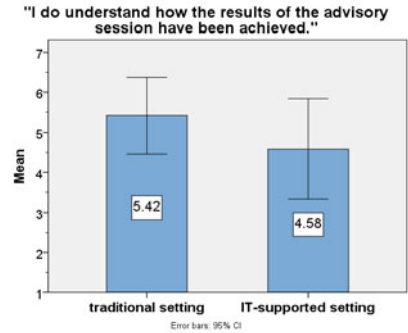


Fig. 3b. Mean Comp2 (error bars: 95% CI)

Overall, seven clients perceived the comprehensibility of the different settings as being equal, while one preferred the traditional setting and another one preferred the IT-supported setting. Five clients and three advisors, however, explicitly stated that the artifact’s visualizations were greatly supporting comprehensibility and transparency. Additionally, one advisor appreciated the navigable process map as an external memory.

Figure 4a and Figure 4b show the means of the clients’ perceived degree of being able to influence the solution finding process (IoAP; avg. TS =5,50 and avg. ITSS = 4,00) and the clients’ perceived possibility to participate in activities (AP; avg. TS = 5,17 and avg. ITSS = 4,50). Again, for both items the average agreement is lower for the IT-supported setting. The results of the two-sided t-test indicate a significant difference between the two means (TS and ITSS) of IoAP ($p < 0.05$, $df = 11$, $t = 2,691$) but no significant results for AP. Hence, our data do not support hypotheses H2.1 and H2.2.

For these results, we obtained the following explanations from the participants. Three clients believed that they could better influence the result of the advisory process in the traditional setting, while one was stating the opposite. Seven participants found that the traditional advisory was more personal than the IT-supported setting, some of them stating their impression that the advisor was focusing too much on the

system. This argument was also raised by advisors, which found it difficult to maintain the conversation with the client while interacting with the artifact.

One client perceived the process in the traditional setting as being more flexible than in the IT-supported setting. Even though advisors were not obliged to overly use the system in the IT-supported setting (or use its functionalities in a specific order), they also found the system to restrict their performance.

Four clients perceived a higher possibility to influence and shape the activities in the traditional setting, three in the IT-supported setting and three claimed to have had the same possibilities in both settings. Only one client seemed to have realized that the order of activities could be changed by skipping activities or selecting previous process steps.

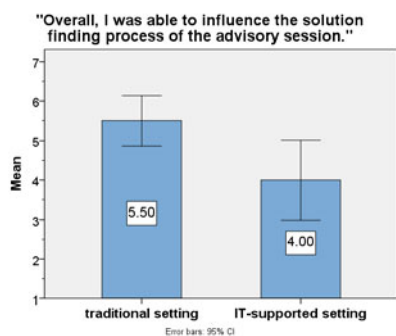


Fig. 4a. Mean IoAP (error bars: 95% CI)

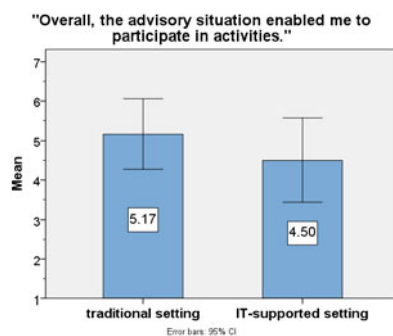


Fig. 4b. Mean AP (error bars: 95% CI)

While five clients would have liked to interact with the system more often, five clients found their interaction with the system to be sufficient. Only one client, however, thought that he would disturb the advisor by interacting with the system. Interestingly, advisors had no reservations in letting the client interact with “their” system, perceiving the clients’ activation as a benefit.

On average, satisfaction with the advisory situation (Figure 5) again was rated higher for the traditional setting than for the IT-supported setting (avg. TS = 5,37; avg. ITSS = 4,90), but the t-test reveals no significant difference. Thus, our data do not support H3.

In the interviews, six clients stated that they would prefer an IT-supported advisory in their next advisory session. Five participants preferred to have their next advisory encounter in a traditional session. Two clients and one advisor proposed to combine the strengths of both sessions (starting the advisory session with traditional face-to-face conversation and consulting the IT artifact only later), highlighting the importance to maintain the advisor-client conversation.

Overall, eight clients would have recommended the IT-supported advisory to others, while five would have recommended the traditional advisory. Three of four advisors enjoyed using the IT artifact and were looking forward to have such tools at their disposal. The remaining advisor, however, felt insecure about using IT with his clients.

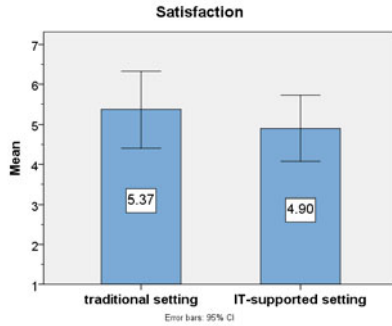


Fig. 5. Mean satisfaction (error bars: 95% CI)

5 Discussion

Looking at the results, we can safely assert that our design for process transparency was not successful in improving the client's overall experience, with all investigated dimensions being rated lower for the new advisory setting. This, however, does not necessarily mean that the design per se was a failure. To the contrary, the clients' ratings of process transparency (comprehensibility of the activities and their results) were rather positive, though leaving room for improvement when being compared to the traditional setting. We suggest that the chosen design for process transparency might be ambiguous, thereby affecting and conflicting with other factors of advisory. For this conclusion we find various indications, which can be related to the system design as well as the introduction of technology into so far predominantly social situations.

Perceived authority and determinism: Though the implemented process structure was consistent with the traditional advisory's course of activities (to which advisors also complied in the traditional advisory scenario), both the advisors and clients found the process depiction to constrain their interaction. As such, the system was perceived to be authoritative and deterministic, imposing its process structure upon the users and restricting the user's control of the process. One client, for example, pointed out that the system could not provide a solution featuring a small amount of shares, whereas in the traditional setting the advisor just "took a note", thereby reassuring the client that he would bear that in mind. The process *map* – that we intended to provide basic orientation and structure – turned out to be conceived as a set of *scripts*, i.e., detailed instructions of how to interact and collaborate to solve the given tasks. Interestingly, when using the system, both the advisors and the clients seemed to feel obliged to stick to the process structure, scarcely changing the order of activities or omitting/revisiting them. Though this allowed the client to associate activities to clearly defined process steps and making their succession transparent, the overall interaction and control was unsatisfying.

Collaboration in advisory: Conceptualizing the interaction of advisor and client as collaborative work (in that they jointly solve the client's problem, being in need of

each others' input to achieve their goals), we borrowed our mechanisms to establish process transparency from CSCW and CSCL research. We believe that in complex advisory situations like investment advisory, learning is an adamant cornerstone of a client's understanding and comprehensibility. In this context, however, the aspect of cooperation deserves some more attention. Research on CSCW and CSCL commonly assumes that work groups are pursuing a common and shared goal. In such work contexts, conflicts between participants are mostly related as *how* to achieve a goal, rather than related to *what* the goal is. In financial advisory encounters, nevertheless, the opposite might be true. The specific results of the service encounter – emerging from the information exchange between advisor and client – are difficult to anticipate before the actual consultation. Additionally, financial products are not only virtual but also credence goods, i.e., clients may not be able to demonstrate whether the success or failure of a purchased product is due to the counseling process, the advisor's (or bank's) efforts or is simply a product of chance. Thus, for the client the journey might actually be the reward – as the result may not easily (or objectively) evaluated, the process leading to a particular result will be closely scrutinized. As discussed above, clients being advised using the prototype system felt that it restricted their interaction with the advisor and concluded that *what they saw was all they could get*. The visualization of the process seemed to confine their problem space to the type of problems the system could tackle and, partly as a consequence, the advisor's solution space to a set of solutions for predefined problems.

Another influencing factor regarding cooperation can be found in the implicit changes of the participants' roles in the technologically supported setting. While the roles of advisor and client and their related (role) scripts [23] seemed to be clear in the traditional setting, the new, collaborative situation implicitly changed the existing role models. Using the cooperative artifact demanded stronger and more active inclusion of the client to accomplish the given activities. To the contrary, the advisors' role tended to change to “coach” the customer in finding the solution rather than providing it themselves.

Focus on the shared artifact: Contrary to the results of our survey, where clients voiced their distrust in banks and their advisors [1], the clients participating in our evaluation found the advisors to be likeable and empathic. The cooperative artifact with shared information spaces clearly shifted the users' attention from interpersonal communication to operating the system. While some clients did not like the idea of the advisor paying more attention to the system than to them, others appreciated the possibility to reflect on the visualizations and interrelations of the financial concepts without the advisor “constantly talking” to them.

In some activities, both advisor and client had problems in correctly interpreting the visualizations and expected interaction, leading to communication breakdowns. The advisor's misinterpretations may be attributed to the lack of training they received in using the system. For the client, however, not every breakdown was problematic – often, the visualization would raise further and more detailed questions. All in all, we speculate that the system's presence clearly interfered with the advisory situation, especially with the client-advisor interaction.

6 Lessons Learned

Establishing process transparency in advisory encounters seems to be a double-edged sword – though possibly allowing for more comprehensibility and understanding of the process flow and its activities, disclosing the underlying mechanisms of advisory can negatively influence the clients' perception of controllability. In this paper, we introduced a design for process transparency that demonstrated such an effect. We argue, however, that the design's failure to consider such an influence provides several lessons to learn. In the following, we will discuss such lessons and show how we incorporated them into our next design iteration.

The depiction of the advisory process as a static, fixed-positioned map has proven to be a poor design choice. In contrast to cooperative work or learning efforts, which might profit from explicit representational guidance to negotiate and relate to a common course of action, in advisory encounters such process representations may lead to a perception of constrained control. In our follow-up design, we therefore abandoned the fixed-positioned process depiction in favor of an implicit illustration of activities and their interrelations. While the system still provides an overview of the process (regarding the basic phases of advisory without implying an order) that may be displayed if required, there are no fixed activities attached. The advisor may use all of the system's functionality in each process phase, whereby interrelations between the functionalities are visualized to support the client's comprehensibility and understanding. If advisor and client, for example, discuss an investment strategy using the system, only relevant information from other activities is displayed – thereby, the client learns about the impacts of activities and the decisions made therein, while the advisor is prevented from "skipping" relevant activities. Such a design of contextualized access to activities makes their relations transparent without forcing actors to use them.

For the users, visualizing the advisory process on the shared technological artifact seemed to imply that all activities have to be accomplished using the system. This clearly is not desirable, as specific activities such as discussing the client's personal needs and wishes or financial goals are better left to the face-to-face dialogue of the client and advisor. At some point of time, however, especially with increasing complexity, directing the dialogue from the inter-personal level to a more focused technology-mediated discussion might be helpful or even necessary (e.g., when contextualizing the client's financial goals with the projected performance of her investment portfolio). Such a situational use of the supporting artifact for specific activities requires smooth transitions from face-to-face conversation to focused interaction with the shared information space and vice versa. Such a requirement cannot be fulfilled by technology itself – therefore, we plan to accompany our follow-up design with practical recommendations for advisors of how to integrate the artifact into advisory while maintaining personal interaction and preventing too much focus on the artifact.

Finally, with all the functionality possibly being incorporated into a software artifact, process transparency design should also make clear that the artifact is not the advisor's replacement. The software system cannot (and, as discussed above, *should not*) account for all possible activities or special cases that might emerge in advisory encounters. Missing functionality or the system failing to meet a client's specific desires, however, must not lead to communication breakdowns. At this, we can learn from the advisors' present working practice – if the advisor is not prepared to

immediately answer a question or fulfill specific demands, he will take a note and inform the client that he will take care of it later; the client is not expecting the advisor to be all-knowing – and she certainly should not expect it from the system.

7 Conclusion

In this paper we have discussed possible designs for process transparency in financial advisory encounters to increase the client's comprehensibility and understanding of the advisory process, its activities and its results. Drawing from CSCW and CSCL research, we used a fixed-positioned, navigable process map to support process orientation and highlight interrelations of activities. Though the comprehensibility and perceived transparency in using the prototype system are rated about as positive as for the traditional setting, our design failed to improve the client's overall experience. We argue that our design may indeed support transparency in advisory encounters, but may also have negative effects on other seemingly important dimensions, namely the perceived controllability and comprehension of the advisory process as well as the interpersonal relationship between advisor and client.

Acknowledgments. The research discussed in this paper is co-financed by the Swiss federal innovation promotion agency CTI.

References

1. Mogicato, R., Schwabe, G., Nussbaumer, P., Stehli, E., Eberhard, M.: *Beratungsqualität in Banken*. Solution Providers AG, Dübendorf (2009)
2. Novak, J.: *Mine, yours...ours? Designing for Principal-Agent Collaboration in Interactive Value Creation*. In: *Proceedings of Wirtschaftsinformatik 2009*, Wien (2009)
3. Buhl, H.U., Kaiser, M.: *Herausforderungen und Gestaltungschancen aufgrund von MiFID und EU-Vermittlerrichtlinie*. *Zeitschrift für Bankrecht und Bankwirtschaft* 20, 43–51 (2008)
4. *Stiftung Warentest: Die Blamage geht weiter*. *Finanztest* 8, 25–30 (2010)
5. Klöckner, B.W.: *Berater test. Peinliche Ergebnisse - große Chancen*. *Bankmagazin* 56, 42–43 (2007)
6. Evers, J., Krüger, U., Reifner, U.: *Beratungsqualität in Finanzdienstleistungen*. Nomos-Verl.-Ges., Baden-Baden (2000)
7. Golec, J.H.: *Empirical Tests of a Principal-Agent Model of the Investor-Investment Advisor Relationship*. *The Journal of Financial and Quantitative Analysis* 27, 81–95 (1992)
8. Eisenhardt, K.M.: *Agency Theory: An Assessment and Review*. *The Academy of Management Review* 14, 57–74 (1989)
9. Nussbaumer, P., Schwabe, G.: *Gemeinsam statt einsam: Kooperative Bankberatung*. *Mensch & Computer* 2010, Duisburg (2010)
10. Schmidt-Rauch, S., Nussbaumer, P.: *Putting Value Co-Creation into Practice: A Case for Advisory Support*. Under Review (2010)
11. Schmidt, K., Bannon, L.: *Taking CSCW seriously*. In: *Computer Supported Cooperative Work (CSCW)*, vol. 1, pp. 7–40 (1992)

12. Kienle, A.: Integration of knowledge management and collaborative learning by technical supported communication processes. *Education and Information Technologies* 11, 161–185 (2006)
13. Guzdial, M., Turns, J.: Effective Discussion through a Computer-Mediated Anchored Forum. *The Journal of the Learning Sciences* 9, 437–469 (2000)
14. Engeström, Y.: Activity theory and individual and social transformation. In: *Perspectives on Activity Theory*, pp. 19–38 (1999)
15. Larkin, J.H., Simon, H.A.: Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science* 11, 65–100 (1987)
16. Zhang, J.: The nature of external representations in problem solving. *Cognitive Science* 21, 179–217 (1997)
17. Suthers, D.D., Hundhausen, C.D.: The effects of representation on students' elaborations in collaborative inquiry. In: *Proceedings of the Conference on CSCL: Foundations for a CSCL Community*, pp. 472–480 (2002)
18. Kienle, A.: *Computerunterstützung für die Organisation menschlicher Kommunikationsprozesse*. Habilitationsschrift, FernUniversität Hagen (2009)
19. Wang, W., Haake, J.M., Rubart, J., Tietze, D.A.: Hypermedia-based support for cooperative learning of process knowledge. *Journal of Network and Computer Applications* 23, 357–379 (2000)
20. Carell, A., Herrmann, T., Kienle, A., Menold, N.: Improving the coordination of collaborative learning with process models. In: *Proceedings of the 2005 Conference on CSCL*, pp. 18–27 (2005)
21. Dillenbourg, P.: Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In: Kirschner, P. (ed.) *Three worlds of CSCL. Can we support CSCL?*, pp. 61–91. Open University of The Netherlands, Heerlen (2002)
22. Briggs, R., Reinig, B.A., De Vreede, G.: The Yield Shift Theory of Satisfaction and Its Application to the IS/IT Domain. *Journal of the Association for Information Systems* 9, 267–293 (2008)
23. Solomon, M.R., Surprenant, C., Czepiel, J.A., Gutman, E.G.: A Role Theory Perspective on Dyadic Interactions: The Service Encounter. *The Journal of Marketing* 49, 99–111 (1985)

A Framework for Supporting Joint Interpersonal Attention in Distributed Groups

Jeremy Birnholtz¹, Johnathon Schultz¹, Matthew Lepage¹, and Carl Gutwin²

¹ Departments of Communication and Information Science
Cornell University, 336 Kennedy Hall, Ithaca NY, USA

² Department of Computer Science, University of Saskatchewan
110 Science Place, Saskatoon, SK, S7N 5C9, Canada

{jpb277, jts228, mc182}@cornell.edu, carl.gutwin@usask.ca

Abstract. Informal interactions are a key element of workgroup communication, but have proven difficult to support in distributed groups. One reason for this is that existing systems have focused either on novel means for gathering information about the availability or activity of others, or on allowing people to display their activities to others. There has not been sufficient focus on the interplay between these activities. This interplay is important, however, because mutual awareness and attention are the mechanisms by which people negotiate the start of conversations. In this paper, we present the OpenMessenger Framework, a system and design framework rooted in the assumption that individual behaviors occur in anticipation of and/or in response to the behavior of others. We describe both the system architecture, and specific examples of the novel implementations it enables. These include techniques for coupling gathering behaviors with display behaviors, and for integrating these into user workspaces via peripheral displays and gaze tracking.

Keywords: Awareness, Attention, Interaction, CMC, CSCW.

1 Introduction

Informal interaction has repeatedly been shown by CSCW researchers to be a key attribute of modern work [2, 37, 47]. Significant efforts over the past 20 years have focused on supporting these interactions via improved awareness of others' presence and activities [21, 31], by improving people's ability to interrupt at appropriate times [19], and by strategically displaying impending interruptions (e.g., [33, 38]). While this work has yielded many research prototypes, the most common tools used in everyday, real-world collaboration still offer only rudimentary support for initiating and concluding informal interactions [9, 46].

One key reason for this is the persistent difficulty of supporting fluid transitions between passive awareness of one's surroundings and engaged attention to a particular person or object. Many have demonstrated the ability of people in face-to-face environments to track others' activities in their visual or auditory periphery, and adjust their own activities or shift their focus of attention accordingly (e.g., [28, 45]).

These behaviors have proven difficult to support online, however [31, 46]. Nearly ten years ago, Schmidt [44] discussed the problems of framing awareness either as an abstract sense of others that is independent of attentional focus *or* as the object of attention itself. He noted that people are constantly monitoring their environment, but also sometimes focusing on others and adjusting behavior accordingly. Since Schmidt's discussion of these issues, however, few systems or frameworks have addressed this dual nature of awareness.

We argue that one way to address this issue is by considering it as a problem of joint attention management, drawing on Clark's [17] notion of *joint activity*. Joint activity occurs when two or more people act individually but coordinate their actions toward achievement of a shared goal. In assessing availability and initiating conversation, we can think of individual actions in terms of *gathering* (Schmidt [44] calls this "monitoring") information about others' availability and signaling, or *displaying* availability for or interest in interaction.

Suppose Alex and Bill work across the room from each other. If Alex moves closer to Bill to *gather* information about Bill's availability, Bill may notice Alex's presence and glance at him. Alex then notices Bill glance and returns it, or, if Bill seems busy, Alex may decide to come back another time. In this way, Alex's approach simultaneously serves as *gathering* and *display*. Closer proximity means that Alex can gather more information, and makes Alex's presence more noticeable to Bill. This triggers Bill's glance, which allows Bill to see that Alex is approaching, and to *display* by glancing at Alex that Bill has noticed Alex's approach.

In contrast to this joint perspective, most existing systems have focused *either* on gathering (e.g., deciding when to interrupt [19, 40]) *or* display (e.g., notifying others of one's status [14, 38]). This separation makes it hard to understand or support the interactive aspects of paying attention via acts of gathering and display. In this paper, we introduce the OpenMessenger Framework (OMF), a flexible and extensible framework for developing joint-attention systems. Through OMF, we define and implement structures to address three problems: 1) discerning the user's focus of attention, and treating this differently when focus is on another person; 2) allowing for easy joint action both during and prior to conversational interaction; and 3) allowing for easy and natural awareness of other users' presence and behavior.

2 Background

When people are aware of or interact with others, we consider them to be managing their attention, which can vary both in intensity and focus, to activities and stimuli in the environment. As such, we define attention more broadly than a purely cognitivist interpretation that considers only a single point of focus [35]. Broadening this definition allows us to account for transitions between more and less active attentional states, and is consistent with evidence from neuroscience on the mechanisms by which people sense and attend to the presence of others (e.g., [41]).

At the same time, our treatment leaves aside debate on the relationship between attention and awareness, as discussed by Schmidt [44]. As Schmidt ultimately points out, however, the interesting practical questions in this domain are not about the underlying awareness and attention mechanisms, but rather the roles played by specific

types of information in affecting behavior. We take as our baseline assumption that abstract awareness in the sense discussed by Dourish and Bly [22] often leads to conscious and focused attention, and that the details of this transition from awareness to engagement are not well supported by today's tools.

We look at these processes within the context of informal interaction and conversation initiation. This is not an exhaustive treatment, but represents a first order approximation [1] of a social problem; one that will enable a foothold in this key area.

2.1 Gathering, Display and Coupling

Attentional behaviors can be usefully framed using the terms *gathering* and *display*. *Gathering* refers to one person getting (either via passive monitoring or active seeking) information about what another is attending to, such as tasks, other people or the gatherer herself. *Display* refers to information about the focus of one's attention that is available for others to gather. We separate these concepts for discussion, but note that they are highly interdependent.

These cues are used reciprocally; initiating interaction is a process of negotiating joint attention via sequences of actions informed by present and prior actions of others [10, 28, 36]. As such, one framework that can be useful in understanding this process is Clark's framework of joint action. This framework allows for the situation of Schmidt's observations about both passive and active states of awareness within the context of behavior aimed at a shared objective or goal. Joint action occurs when people act in the belief that they are part of a collective activity, in which their actions occur in response to coordination signals from another party. We argue that gathering and display are what Clark [17] calls component moves in joint action.

Consider again a case in which Alex wants to talk to Bill. Alex must get Bill's attention, which he does using both *gathering* and *display*. Glancing quickly allows Alex to *gather* information about Bill's likely availability; he then walks toward Bill, which serves to *display* Alex's intent to start a conversation. Bill then can respond by using gaze and body position to *display* his own interest.

These individual actions can be interpreted within the framework of joint activity. If Bill is wearing headphones and does not look up as Alex approaches, for example, that could signal either that Bill is unaware of Alex, meaning that Alex needs to get closer and louder; or that Bill is aware that Alex is trying to talk to him, but signaling that he (Bill) is unavailable. This judgment on Alex's part is based on information about Bill, the outcomes of recent actions, and knowledge of the context.

One key attribute of Alex's approach for the purposes of attention management is the relationship between gathering and display. Alex's movement toward Bill to gather information necessarily functions simultaneously as an act of display, because Bill can see him approaching [10]. We can therefore say that these instances of gathering and display are tightly coupled. In face-to-face interactions, people's ability to notice others (e.g., [41]) relies on this coupling. We perceive others' gathering because it is visible in ways that we can attend to.

Returning to Clark's terms, we can define coupling as the extent to which a particular action is visible and noticeable to others. Table 1 illustrates coupling relationships between gathering and display. In face-to-face approaches (top left cell), physical proximity and eye gaze are effective ways to display attention because

gathering and display are tightly coupled [10]. In contrast, many behaviors that are tightly coupled in face-to-face interactions have different relationships online [5], as illustrated in the three remaining cells of Table 1. If Alex and Bill are spying covertly on each other via webcam, for example, gathering and display may be completely decoupled (bottom right cell). Alex’s gathering is not displayed at all, and vice versa.

Table 1. Coupling of Gathering and Display

		Alex Gathers	
		Displayed	Not Displayed
Bill Gathers	Displayed	Face-to-Face Approach	Some IM conversations
	Not Displayed	Some IM conversations	Spying, covert looking

This helps to address a key question identified by Schmidt [44]: how do people regulate the obtrusiveness of their behavior? In face-to-face encounters, obtrusiveness is often a function of the salience of display. Staring at somebody’s screen or standing very close to them, for example, are very salient forms of display that are also obtrusive [10]. Online, however, it is possible to get detailed information about others without appearing obtrusive or even being visible at all. These examples of asymmetric coupling relationships have a significant effect both on how people regulate their behavior, and on the feasibility of joint action.

One example is the “appear offline” option on instant messaging (IM) clients used by those wishing to avoid interruptions [9]. People using this option can gather information about others on their contact list without those others even knowing that such gathering is possible (because they cannot see that the gatherer is online). This sidesteps the key role that obtrusiveness ordinarily plays in attracting and negotiating attention and has significant consequences for joint action, because people cannot respond to actions of which they are unaware. In other words, it is the coupling of gathering and display in face-to-face interactions that helps facilitate joint action.

2.2 Gathering Is Display; Display Is Gathering

Despite Schmidt’s discussion of awareness as a duality of gathering *and* display, most systems and theoretical frameworks for addressing the problem have focused either on one or the other.

Early media space systems used cameras to provide video views of others in their offices [22, 23, 29, 31]. Cameras, however, were thought by some to be invasive [15, 18]; and the systems did not support the subtleties of negotiating interaction [32]. These problems reflect the de-coupling of gathering and display in that one user could view (i.e., gather) video of another, without a clear display that this was taking place.

These early video experiences led many to experiment with the notion of a “virtual approach” (reviewed in [46]). In our terms, this work can be characterized as an attempt to increase coupling between gathering and display by displaying to an observed party that gathering is taking place, and that conversation may be desired. The idea was that the approach would facilitate interaction more naturally by allowing for multiple levels of gathering, and by displaying activities to the observed parties.

As such, “approaches” often involved replicating the sequence of actions typically involved in initiating face-to-face conversations. Several systems allowed users to, for example, “glance” at others to discern their availability [34], and then follow a series of progressively more informative steps eventually resulting in conversation. At each step, the observed party would have to respond in kind (e.g., with a “glance” of their own) to proceed with the interaction.

More recent systems such as Community Bar [39] allow for a continuum of awareness states. In Community Bar, these must be manually updated via independent “focus” (how much information is seen about others) and “nimbus” (how much information is revealed to others) sliders. These terms come from work by Benford and Fahlen [8] and Rodden [43] that aims to distinguish between being the object of somebody’s attention (their nimbus) and focusing on somebody (one’s focus). While this distinction is a useful one, the Community Bar implementation is problematic in that it renders the coupling relationship between gathering and display dependent on the combined status of independent users’ sliders being manipulated in parallel. That is, Alex could choose to reveal more information to Bill (via a nimbus slider) even as Bill is reducing the amount of information he sees about Alex (via a focus slider).

From a joint action standpoint, designs that foster multiple levels of awareness and sharing are an improvement in that they allow for coordinated activity. These systems were critiqued, however, for requiring lockstep and seemingly artificial sequences of behavior. By this, we mean that real-world gathering behaviors (e.g., glancing, walking up to somebody) are easily noticed and responded to because, unlike pointing at buttons in an interface or watching a PC window for notification of an incoming virtual “glance,” they are the actions that naturally occur to assess availability, respond to somebody to avoid appearing rude, and/or to start a conversation [28, 36].

We argue that effective support for a joint action approach to attention requires consideration of the interplay between gathering and display. Specifically: 1) acts of gathering must be coupled to displays or notifications, and 2) these displays must be easily noticed and responded to via subsequent acts of gathering, that must also be displayed. Gathering must be displayed and displays must be gatherable, *ad infinitum*.

2.3 Approaching Is Interacting

A second key question in this area concerns the specific information that causes people to adapt their behavior in response to the actions or reactions of others. There are useful lessons from face to face interaction that can be considered here.

Goffman [30] argues that human behavior around others is performative; it is often intended to convey information or impressions to others. Sudnow [45] discusses the importance of glances in assessing others’ behavior and availability. He notes that people in public settings know that others may glance at them and act accordingly, such as by putting on headphones or adjusting posture to appear busy [10]. Sudnow [45] refers to these as “glanceable states,” in that status can be discerned via a glance.

In some ways, the glanceable state is reflected in current interaction tools, such as the IM contact list. The intent of the list is to concisely summarize who is online and available. There has also been work aimed at improving this information by automatically updating status information (i.e., availability) via sensors (e.g., [7, 25]).

One problem with the IM contact list, however, is that recent work has focused *either* on the problem of interruptions [40] and developing systems that allow for better timing of interruptions (i.e., better gathering of information; [4, 19]) *or* on techniques for unobtrusive notification of impending interruption (i.e., better display; [6, 14, 38]). Considering these problems in isolation ignores a key component of our joint action argument: acts of gathering and display occur in response to each other. Approaching somebody does not occur prior to interacting; it is part of the joint action of initiating conversation. That is, approaching *is* interacting.

People do not respond the same way to interruptions from all others, such as work collaborators vs. social friends, [20], and also may behave differently when they know their behavior is being monitored by others [27]. Moreover, not all interactions result from interruptions; many result from serendipitous mutual attention [36]. Thus, our second argument is that each act of gathering and display, however preliminary from the standpoint of starting verbal conversation, must be considered as component behaviors in an interaction (or joint action) to which others should be able to respond.

A real-world approach progresses from distant observation – characterized by less detailed gathering and less salient display – to closer observation – characterized by more detailed gathering and more salient display due to physical closeness. We advocate a similar progression online, consisting of multiple types of interactive behavior that enable both gathering and display to take place. As with the face-to-face approach, we emphasize that what is important is not reciprocal instances of identical behavior (i.e., a glance must be followed by a glance), but rather a general correlation structure between behaviors such that the amount of detail that can be gleaned from a particular gathering behavior roughly correlates with the salience of the display behavior with which that gathering is coupled.

3 The OpenMessenger Framework

Supporting a joint-action approach to attention management requires mechanisms for coupling gathering and display behaviors, and for treating these as interaction. In this section of the paper, we present the OpenMessenger¹ Framework (OMF), a software framework and application for addressing these issues. We aim to make two contributions: 1) an extensible software framework for experimental exploration of issues related to awareness, and 2) an implementation example with novel gathering and display mechanisms.

3.1 Supporting Gathering and Display at the Framework Level

To effectively support joint action in attention management, we need a conceptual architecture that supports gathering and display of information, and the coupling of these behaviors to each other. This is accomplished in OMF with abstractions called *sensor managers*, *monitors*, *awareness events*, and *views*. Data about user activities is captured from hardware sensors by *sensor managers*, and is then analyzed and distributed to other OM clients by *monitors*. Transmission of the data is via *awareness*

¹ Note that the word “open” in OMF refers to open-plan offices, which were our inspiration for this work. We will happily share OMF source code, but “open” does not imply open-source.

events passed from clients to the OM server, and then to all clients in a pre-defined group (see Figure 1). Event information is made perceptually available to users in a view that could be a visual, auditory, or tactile display.

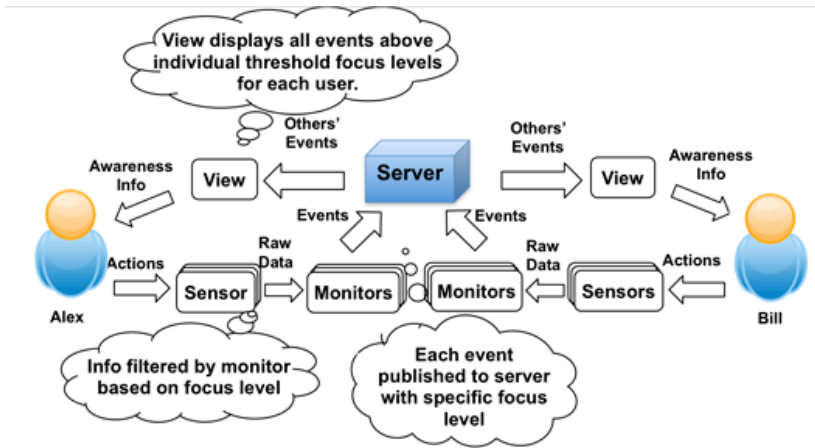


Fig. 1. Information flow in the OpenMessenger Framework

3.1.1 Sensor Managers

Gathering information about others is supported by *sensor managers*, which are C# objects on the client PC that regularly sample data sources in the environment such as hardware devices or operating system information. Sensor managers provide information to other components using a publish/subscribe model. Subtypes of the Sensor-Manager class are singletons.

3.1.2 Monitors

To allow people to gather useful information about others while still restricting the flow of potentially sensitive sensor data (as in [7, 24]), and avoiding user inundation with information, OMF uses *monitors* to distill raw sensor data. OMF monitors are C# objects on the client that subscribe to one or more sensor managers and process the raw sensor data for publishing to the server as *awareness events*. These events are sent to the server, which broadcasts them to all connected clients (including the sender, though the sender does not use them). Windows Communication Framework (WCF) is used for network communication (based on TCP sockets). When an awareness event is received, the server stores it in the server *event cache*. With this cache, event data can be immediately sent to newly connected clients.

As an example, the monitor for a keyboard activity sensor might release keystroke frequency, but not reveal which keys were pressed. This allows for conveying information about activities without releasing potentially sensitive data, and also takes data that may not be meaningful in raw form (e.g., sound level) and processes it to provide useful information (e.g., presence of sound above a conversational threshold).

Importantly, monitors are abstractions of sensor managers. While raw sensor data generally comes only in one form, data can be used in many ways. For example, a

microphone sensor could indicate sound above a threshold, while another monitor could use speech recognition on the same data to determine if there is a conversation going on. As with sensor managers, subtypes of the Monitor class are singletons.

3.1.3 Social Monitors

Monitors, as described so far, operate similarly to previous systems when the user is focused on objects in the task or environment [26]. Our joint-action perspective, however, requires a way for users to dynamically detect and respond to attention from other users. Thus, a unique feature of OMF is special support for situations when the user's focus of attention is on another user. This is accomplished via *social monitors*, a sub-type of the monitor class that communicates this information.

Social monitors are necessary because attention to another person is distinct from attention to a task element or interface object. As discussed above, attention to people is interactive; it often occurs often in expectation of and/or in response to another act of gathering or display. From an implementation standpoint, social monitors are unique in that they affect the intensity of attention that one user is paying to another, which we define later as *focus level*. Social monitors determine when one user is attending (e.g., via mouse or other input data) to another user's representation (e.g., an on-screen avatar) on a view. This is accomplished via data about attentional focus (e.g., an x,y coordinate pair from the mouse or eye tracker monitor) and knowledge about the arrangement of avatars on the view. When one user is determined to be attending to another, focus levels are adjusted accordingly (see below).

Table 2. Example monitors and associated sensor managers in OMF

Monitor	SensorManager Used (API)	Analysis
<ul style="list-style-type: none"> • Typing Activity • Keypresses 	Keyboard (DirectInput)	<ul style="list-style-type: none"> • # keys pressed since last event • Record keys since last event
<ul style="list-style-type: none"> • Screen Activity • Current Window • Screen Contents 	Screen (Windows API)	<ul style="list-style-type: none"> • Compare successive screenshot frames • Record title of current focus window • Record screen image
<ul style="list-style-type: none"> • Sound Presence • Sound Level • Number of Speakers 	Microphone (DirectSound)	<ul style="list-style-type: none"> • Compare microphone level to thresholds • Track level over time • Analyze sound for voices
<ul style="list-style-type: none"> • Overall Presence 	Keyboard, Screen, Microphone	<ul style="list-style-type: none"> • Time since last sensor change
<ul style="list-style-type: none"> • Visual Presence • Visual Changes • Number of People 	Webcam (DirectX)	<ul style="list-style-type: none"> • Compare successive frames for changes • Analyze frame for objects (e.g., people) • Record webcam image
<ul style="list-style-type: none"> • Mouse Focus Target 	Mouse (UI toolkit)	<ul style="list-style-type: none"> • Determine avatar mouse is pointing at
<ul style="list-style-type: none"> • Eye Focus Target 	Eye Tracker (Custom API)	<ul style="list-style-type: none"> • Determine avatar the user is looking at
<ul style="list-style-type: none"> • In Conversation 	Chat Window (OMF events)	<ul style="list-style-type: none"> • Determine who the user is conversing with

3.1.4 Sensor and Monitor Examples

Example OMF sensor managers and monitors are shown in Table 1. This is not an exhaustive list of possibilities; most of these specific examples were chosen based on our own experience with the first OM application and based on prior work on

predicting availability with sensor data [4, 7, 26]. That work suggests that keyboard activity, idle time, and the presence of sound can be very helpful in determining availability. Gathering active window titles and screen snapshots are sources that were used in the initial version of OM, but were seen by some users as too invasive [11]. We therefore added a monitor for this data source that only detects screen changes, rather than transmitting the actual screen image. The eye tracker and mouse monitors are unique to the OMF in that they are social monitors, described below.

OMF also provides a simple extension mechanism for adding new monitors for additional analyses, or new sensor managers for new data sources. This requires some coding by the application programmer – creating a new subtype of Monitor or SensorManager, and linking APIs for new hardware sensors to these classes – but the process is simple and existing classes can be used as templates.

3.1.5 Views

Gathering and display behaviors are rooted in the perceptual availability of awareness information. Gathering cannot occur if this information cannot be perceived, and display has no purpose if the information is unavailable to others. Awareness information in OMF is made perceptually available via a *view*. OMF provides a basic visual view based on earlier OM systems [11]. However, application programmers can easily develop new views within the .NET environment (e.g., using XAML or Windows Forms), or can build displays that use sound or other feedback mechanisms. Views are client-specific (i.e., they are not WYSIWIS), and multiple views can run simultaneously on the same client. A standard data structure for each user simplifies the storage of view-based information such as an avatar image, user name and ID, and the most recent awareness event data received pertaining to that user.

3.2 Focus Level: Degree of Attention and Coupling

As noted above, people do not respond the same way to interruptions from different people, and some interruptions result from serendipitous mutual attention. Alex may be available to Bill but not to Cathy, for example. Support for variation in treatment of contacts, and for awareness of mutual attention requires the capacity to share different information with certain contacts. Moreover, as Alex receives attention from Bill, he should be able to easily respond by displaying his own attention to Bill (indicating interest or availability) or to a task or another user (indicating that he is busy). In OMF, this is accomplished via the *focus level* mechanism, which represents the level of attention that one user is paying to another.

It is through focus levels that acts of gathering are coupled to acts of display, and that the correlation structure of gathering and display behaviors is maintained. First, the focus level of one user on another determines what information the first can gather about the second. Each awareness event includes a threshold focus level for event data display. This means that if Alex releases an awareness event with threshold level X , Bill will only see this event if his focus level on Alex is greater than or equal to X .

More detailed information is therefore assigned a higher focus level.

Second, the focus level of one user on another is used to notify the observed user (i.e., display to them) that gathering is taking place. Notification increases in salience as focus level increases. That is, Alex receives increasingly salient notifications as

Bill's focus on him increases. Our OMNI system described below uses visual salience, but this could be accomplished with sounds or other cues as well. The key is that more detailed gathering behaviors correlate with more salient displays.

Focus level is represented in the system as an integer (range: 0-5, with 0 as minimum) for each user's level of focus on each other user. Focus level is updated dynamically via social monitors, which detect one user's focus on another. The 0-5 range is based on experience with the initial OM version, suggesting that six levels provides enough variance for multiple means of gathering, and for displaying varying levels of interest without distracting the user. Further, focus is uni-directional. That is, Bill's focus on Alex does not vary directly with Alex's focus on Bill. Rather, each client stores a list of the current client's focus level on all other users. As such there are two discrete focus level variables that describe Alex and Bill's focus on each other. Thus, OMF is not a strict reciprocity-based system [46].

Information displayed at each focus level in the current OM is listed in Table 3. This sequence is preliminary, but based on our experience with sharing similar information in previous OM versions. The focus level mechanism could also be exploited in more powerful ways by providing multiple information streams at each level, or by providing progressively more detailed information from a single sensor.

Table 3. Example focus levels and corresponding monitors. Note that information display is cumulative, such that each level includes the lower levels as well.

Focus Level	Data/Monitor (see Table 1)
0	Overall Presence
1	Keyboard Activity, Idle Time
2	Screen Activity
3	Current Window Title
4	Microphone Activity
5	Current Gaze Focus

3.2.1 Changing Focus Level

Focus level is increased via explicit indicators of attention from social monitors (e.g., mouse and eye-gaze data) and decreased over time via attenuation. With the mouse social monitor, for example, if Alex's cursor hovers over Bill's avatar for a predetermined length of time the social monitor produces a *focus change event* that is distributed to all clients.

Testing of earlier OM versions [11, 13] suggests that these trigger mechanisms must be carefully designed. For example, users tended not to use OM's awareness features if it took more than a second for information to appear. We therefore increase focus level after one second of sustained attention to an avatar, and subsequently increment focus level by one after each second of continued attention. This continues until the observer stops focusing on the target, or focus level reaches its maximum. Focus then decreases over time at a linear rate when the observer stops focusing on the target. Decrease is slower than increase: 15 seconds to decrease from level 5 to 0, compared with 5 seconds to increase from level 0 to 5.

As with views, the mechanisms for changing focus levels can be extended in OMF. For example, it would be simple for an application programmer to add a social

monitor that changes focus levels through explicit keystrokes or commands, or that changed the way that focus level increases and decreases.

3.2.2 Event Filtering

As noted above, monitors generate events to be displayed only to users with at least a threshold focus level on the user generating the event. This presents the problem of how to broadcast events: should they be shared with all users, but displayed only to those viewing at the appropriate focus level (receiver-filter), or shared only with users at a particular focus level (sender-filter)? We chose the receiver-filter strategy to reduce the complexity of the system in terms of server query overhead. Given relatively low numbers of connected clients and sensors, broadcasting the data to all clients at once involves fewer operations and is simpler than a query-based model. This might be reconsidered, however, in an environment with a substantially greater number of clients or sensors. This means more network traffic, but this is acceptable since overall bandwidth requirements are low.

3.2.3 Text Chat

To support conversation, the OMF provides basic text chat. This is supported by the framework via the MessageEvent, a type of client-generated OM event characterized by a short text message, a sender and a receiver. The chat window implementation is similar to other text-based chatting systems, so we do not provide a detailed description here. However, OMF chat is novel in that conversations are a potential data source – that is, the ‘In Conversation’ monitor keeps track of who is talking to whom (Table 1), and can use this to help determine attention and focus level.

3.4 An OMF Implementation Example: Eye Tracking and OMNI

To illustrate the capabilities of the OMF, we here describe our current implementation of a system called OMNI, which combines a social monitor that reads data from a head-mounted eye tracker, and a projected peripheral-vision awareness view (Figure 3). OMNI is an OMF implementation example intended for information or other office workers who work primarily at a desk but benefit from frequent informal interaction with potentially remote colleagues (e.g., designers [10], engineers, or researchers [2, 37]). Those with a different work environment may benefit from a different OMF implementation. Elements of this system have been reported previously [12], but this is the first explication of OMNI at the system level.

3.4.1 The EyeTracker Social Monitor

One problem that the OMF enables us to address is the difficulty of mapping natural face-to-face behaviors (e.g., glancing) to online interfaces. Joint action requires the capacity for easy response, and one way we accomplish this is by using an eye/head tracker to capture the user’s visual focus. Eye tracking is particularly appropriate for this context because interpersonal attention is often conveyed via the eyes. Looking at somebody implies interest, and looking away implies disinterest [3, 28, 36].



Fig. 2. Screen shots of two OMNI displays in use, with rows denoting what users Alex and Bill see. In Alex’s view, as Alex focuses on Bill, a halo appears around Bill’s avatar along with a line to indicate that he is being focused on. On Bill’s screen Alex moves closer to Bill’s avatar as Alex’s focus level on Bill increases. (A) is the remote user’s avatar, (B) is the local user, and (C) is sensor information about the remote user.

We use an ASL H6 head-mounted eye tracker and a Flock of Birds magnetic head tracker in our current system. We acknowledge this is a cumbersome and expensive device, but note that eye tracking technology is rapidly falling in cost as it becomes possible to track unobtrusively and inexpensively via webcams (e.g., [16]). While our specific implementation necessarily relies on the details of our hardware, the general approach to eye tracking that we describe here is not device-dependent; it could be adapted to any device that provides (x,y) coordinate pairs on a defined plane in space.

The eye tracker’s API libraries allow for the definition of planes in the real-world space. Each eye-tracker data point consists an (x,y) coordinate pair on a particular plane. Augmented by basic knowledge of the size, location, and resolution of the display that are defined as parameters, we determine when a user is visually attending to elements of the display; and in particular, when they are looking at another person’s avatar. When this occurs and fixation is prolonged, the social monitor increases the focus level on the observed party.

3.4.2 The OMNI Awareness View

One problem with existing systems is the difficulty of noticing when others gather information. Where social and peripheral attention processes mean that people constantly monitor their real world surroundings for the presence and gaze of others [41], limited screen space and window occlusion can limit attention to avatars or notifications online [38], and these displays can be distracting as well [12].

The OpenMessenger Notification and Interaction (OMNI) view is intended to exploit the properties of human visual perception in concert with the eye tracker social monitor and OMF. In the OMF, OMNI is classified as a *view*. The OMNI display consists of avatars representing contacts currently logged in, in addition to basic

awareness information about these contacts. The display is projected onto a surface behind and above the user's primary monitor (or displayed on a large screen in the same space), so that it appears in the periphery of the user's visual field. Avatars are arranged in a semi-circle around the user's primary monitor, with their spatial arrangement initially determined by the order in which users login to the OM server, but this could be changed by the user to reflect natural groupings based on external factors such as project roles or relationship types.

OMNI has two primary uses: 1) displaying others' gathering behavior, and 2) facilitating gathering information about others.

Displaying. For displaying others' behavior, two visual parameters are manipulated: physical distance and motion. Physical distance of a contact's avatar from the user's body in the real world (presumed to be at the center of the display, though this could be altered) is inversely related to OM focus level. Contacts with a higher focus level on the current user (i.e., those that are gathering information) appear closer to the user. Distance between a user and a contact with focus level zero is defined by the vertical height of the display (i.e., the maximum possible distance given screen size constraints). Distance decreases by 20% of the overall distance with each increase in focus level until the avatars are touching at focus level five.

In line with findings from [6], we use motion to attract user attention in the event of change. As a contact increases their focus on the current user, the contact's avatar gradually moves toward the user on the OMNI display. Movement rate correlates directly with the rate of increase in focus level and occurs at a linear rate. As focus level decreases, the avatar moves away from the center at a rate slower than the approach. Movement away occurs more slowly to minimize distraction in the event that a contact looks back at an avatar soon after looking away.

Gathering. When the current user wishes to gather information about a contact, they focus on that contact's avatar (using a mouse, eye tracker, or other device with a social monitor). As they maintain focus on this avatar, more monitor information is displayed (see Figure 2) about that contact, reflecting the increase in focus level.

At this point, our social monitor and view architecture becomes quite powerful in supporting joint action and the notion of approach as a form of interaction. Consider the case where Alex wishes to interact with Bill. Alex looks at his OMNI display

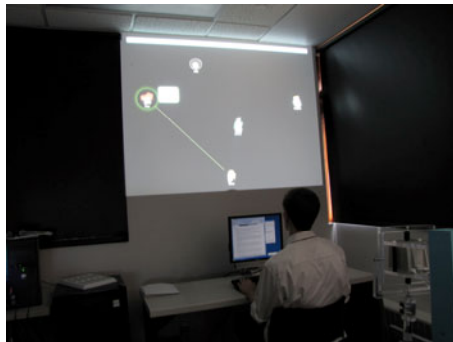


Fig. 3. A user seated in front of an OMNI display, with four remote contacts. The user is focused on the leftmost contact, as indicated by the line.

and sees that Bill is online. He then fixates on Bill's avatar to get more information (Figure 2) about him. As Alex gets more information (and his focus level on Bill increases), Alex's avatar will move toward Bill's on Bill's OMNI display. Bill notices this and looks at Alex's avatar, thus increasing Bill's focus level on Alex. As Alex's avatar moves toward Bill, more information about Alex (from his monitors) will appear next to Alex's avatar on Bill's display. If Bill continues to look at this information, (i.e., as Bill's focus level on Alex increases), Bill's avatar on Alex's display moves toward Alex. Alex knows that Bill has noticed him, and can watch his display to see if Bill continues to attend to him. With sustained attention, they could opt to start a conversation or simply be aware of each other's interest.

4 Discussion

By focusing on both gathering and display of awareness information as interdependent components of the joint activity involved in negotiating mutual attention, OMF provides an extensible and general framework for experimenting with the management of interpersonal attention. It is the product of four years of development efforts, including three versions of OM. Here we discuss the implications of some key design decisions and assess the OMF's technical properties.

4.1 Sensor Data for Both Gathering and Display

We began by noting that prior systems often considered gathering and display in isolation, but that in reality these behaviors often correlate or even overlap. One unique and important attribute of the OMF is that data from sensors are used for both gathering *and* display. This stands in contrast to previous systems, which have largely treated sensor data primarily as a way to gather information about others (e.g., [7, 24]). OMF's sensor/monitor/focus level architecture is agnostic about who is "starting" an interaction or who is "observing" vs. "being observed." It uses sensor data to determine the focus of attention of *both* the observed party *and* the observer.

The implications of this approach are illustrated by the Eye Tracker sensor. If the eye tracker detects Alex looking at Bill's avatar, this single awareness event triggers both an increase in the amount of information Alex sees about Bill (gathering), as well as a notification to Bill that this is taking place (display). In this way, gathering and display are coupled more naturally, and it is technically possible to support the interactive aspects of awareness discussed by Schmidt [44].

More research and experimentation, however, are needed to implement mechanisms for interaction via gathering and display in ways that will not confuse or overwhelm users. This is the focus of our current and future research.

4.2 Approaching, Interacting and Privacy

Our second key point was that approach is a form of interaction, and people should be notified when others gather information about them. While the OMF supports notification for joint action purposes, this does not mean that notification will occur every time information is accessed. From a privacy standpoint, OMF's architecture supports but does not strictly enforce notification. Using Boyle's [15] distinction between

confidentiality and solitude, we think about monitors as protecting the confidentiality of detailed sensor data, while focus levels protect solitude by reducing the amount of information that is displayed at any one time.

The combination of these mechanisms means that there are two potential cases where gathering could occur without notification. The first of these results from possible variation in views between users. It is possible that a user could implement or select a view that, for design or other reasons, does not display instances of certain types of gathering by other users. In these cases, the user would not be notified when these types of gathering took place. Second, receiver-side filtering of awareness events means that data are sent to clients regularly, but notification is provided to an observed user only when the data are accessed via a view. As data is stored on the client, it is theoretically possible that the user could hack the system for access in a way that does not result in notification. While these issues merit consideration, we do not feel they present a major threat to privacy because they seem unlikely to occur often in trusted groups of collaborators, and because the OMF provides other privacy controls that regulate what information is shared. Technical Assessment of OMF.

Generality. The OMF is general in that it can theoretically handle any type of awareness-based attention management in which users gather information about the displayed activity or status of their collaborators, and adjust or update their own activities based on this information. The framework can easily be extended to incorporate data from any software or hardware sensor providing a data stream that can be parsed by an OMF monitor. This could include information to support nonverbal or semi-synchronous communication, or those that provide additional verbal/synchronous channels (e.g., audio, video).

The framework architecture also supports easy development of novel mechanisms for increasing or decreasing focus. While our sample implementation increases focus level when another user's avatar is the focus of attention, this could be extended to increase focus when users are focused on the same object (e.g., a document), or even when certain users are focused on each other.

Flexibility. The ideas behind OMF have been tested primarily with synchronous or semi-synchronous interaction, but the design of the framework is such that the architecture and general implementation approach could be adapted for use in other systems - other awareness / messaging systems, but also in a variety of other applications (e.g., shared editors, workflow systems, or document sharing).

The framework is flexible enough to be used in a variety of contexts, although we have primarily developed and tested it in a desktop/office environment. The general architecture, however, means that a group of users could be using individual clients with very different sensors, sensor managers, monitors and views; but these could all smoothly exchange data via the common structure of awareness events and focus level. In future we plan to adapt the system to mobile devices and sensors.

Performance. The nature of the OMF and current implementations means that the computation and bandwidth requirements for the framework are small. In particular, sensing, monitoring, and distribution of information functions of the OM framework are lightweight. We will consider performance further in future work.

4.3 Limitations and Future Work

There are several limitations that provide opportunities for substantial future work.

One clear limitation in our assessment is the lack of a field or laboratory evaluation, and this is part of our immediate future plans. We do not focus on this here, but instead point to several other key questions. First, we provide an operational and conceptual framework for supporting joint attention management, but leave aside substantial questions about how sensor data should be parsed, interpreted and aggregated as the number of sensors and granularity increase. Parsing sensor data has received substantial attention in recent years (e.g., [7, 25, 42]) and we aim to incorporate and build on these techniques. Relatedly, we currently use a naïve approach to increasing focus levels that assigns equal weight to different modes of input. We aim to develop a more sophisticated model, defined within OMF, that reflects the nuance of cues and modes of social attention.

5 Conclusion

We have presented a joint action approach to interpersonal attention management in supporting awareness and informal interaction. In our approach, actions are assumed to occur in anticipation of or in response to acts by others. Our OpenMessenger Framework provides operational solutions for the problems of: 1) discerning focus of attention, and treating this differently when focus is on another person; 2) allowing for joint action both during and prior to conversational interaction; and 3) allowing for awareness of other users' presence and behavior. The software and examples are available at: <http://collabtech.hci.cornell.edu/projects/openmessenger.php> so that other researchers can further test and explore the idea of joint attention.

Acknowledgements. We thank all who have helped develop OMF, especially Sigurd Teigen, Jeeseung Na, Maryam Mustafa, Suchi Agicha, Oleg Krohkin, and Yang (Leon) Liu. This project is supported in part by the USDA Cooperative State Research, Education and Extension Service (Hatch Project #NYC-131439), the US National Science Foundation (#IIS-0942659), and the Institute for the Social Sciences at Cornell University.

References

1. Ackerman, M.: The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human Computer Interaction* 15(2/3), 181–203 (2000)
2. Allen, T.J.: *Managing the Flow of Technology*. MIT Press, Cambridge (1977)
3. Argyle, M., Cook, M.: *Gaze and Mutual Gaze*. Cambridge Press, Cambridge (1976)
4. Avrahami, D., Fussell, S., Hudson, S.: IM waiting: timing and responsiveness in semi-synchronous communication. In: *Proc. ACM CSCW*, pp. 285–294 (2008)
5. Bailenson, J.N., Beall, A.C., Blascovich, J., Turk, M.: Transformed Social Interaction: Decoupling Representation from Behavior and Form in Collaborative Virtual Environments. *Presence* 13(4), 428–441 (2004)

6. Bartram, L., Ware, C., Calvert, T.: Moticons: detection, distraction and task. *International Journal of Human-Computer Studies (IJHCS)* 58, 513–545 (2003)
7. Begole, J., Matsakis, N., Tang, J.: Lilsys: Inferring Unavailability Using Sensors. In: *Proc. ACM CSCW*, pp. 511–514 (2004)
8. Benford, S., Fahlen, L.: A spatial model of interaction in large virtual environments. In: *Proc. ECSCW*, pp. 109–124 (1993)
9. Birnholtz, J.: Adopt, Adapt, Abandon: Understanding Why Some Young Adults Start, and then Stop, Using Instant Messaging. *Computers in Human Behavior* 26(6), 1427–1433 (2010)
10. Birnholtz, J., Gutwin, C., Hawkey, K.: Privacy in the open: how attention mediates awareness and privacy in open-plan offices. In: *Proc. GROUP 2007*, pp. 51–60 (2007)
11. Birnholtz, J., Gutwin, C., Ramos, G., Watson, M.: OpenMessenger: Gradual Initiation of Interaction for Distributed Workgroups. In: *Proc. ACM CHI*, pp. 1661–1664 (2008)
12. Birnholtz, J., Reynolds, L., Mustafa, M., Luxenberg, E., Gutwin, C.: Awareness Beyond the Desktop: Exploring Attention and Distraction with a Projected Peripheral-Vision Display. In: *Proc. Graphics Interface* (2010)
13. Birnholtz, J., Tang, D.: Sharing Awareness Information Improves Interruption Timing and Social Attraction (in preparation)
14. Booker, J.E., Chewar, C.M., McGrenere, J.: Usability testing of notification interfaces: are we focused on the best metrics? In: *Proc. ACMSE*, pp. 128–133 (2004)
15. Boyle, M., Greenberg, S.: The Language of Privacy: Learning from Video Media Space Analysis and Design. *TOCHI* 12(2), 328–370 (2005)
16. Chau, M., Betke, M.: Real Time Eye Tracking and Blink Detection with USB Cameras. *Boston University Computer Science Technical Report* (2005)
17. Clark, H.H.: *Using language*. Cambridge University Press, New York (1996)
18. Clement, A.: Considering privacy in the development of multi-media communications. In: *Computer Supported Cooperative Work*, vol. 2, pp. 67–88 (1994)
19. Dabbish, L., Kraut, R.: Controlling Interruptions: Awareness displays and social motivation for coordination. In: *Proc. ACM CSCW*, pp. 182–191 (2004)
20. Davis, S., Gutwin, C.: Using Relationship to Control Disclosure in Awareness Servers. In: *Proc. Graphics Interface 2005*, pp. 75–84 (2005)
21. Dourish, P., Belotti, V.: Awareness and Coordination in Shared Workspaces. In: *Proc. Proc. ACM CSCW*, pp. 107–114 (1992)
22. Dourish, P., Bly, S.: Portholes: Supporting awareness in a distributed work group. In: *Proc. ACM CHI*, pp. 541–547 (1992)
23. Fish, R.S., Kraut, R., Chalfonte, B.: The VideoWindow system in informal communication. In: *Proc. ACM CSCW*, pp. 1–11 (1990)
24. Fogarty, J., Au, C., Hudson, S.: Sensing from the Basement: A Feasibility Study of Unobtrusive and Low-Cost Home Activity Recognition. In: *Proc. UIST*, pp. 91–100 (2006)
25. Fogarty, J., Hudson, S.E., Atkeson, C.G., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J.C., Yang, J.: Predicting Human Interruptibility with Sensors. *TOCHI* 12(1), 119–146 (2005)
26. Fogarty, J., Lai, J., Christensen, J.: Presence versus Availability: The Design and Evaluation of a Context-Aware Communication Client. *International Journal of Human-Computer Studies (IJHCS)* 61(3), 299–317 (2004)
27. Forsyth, D.: *Group dynamics*. Brooks/Cole, Pacific Grove (1998)
28. Frischen, A., Bayless, A.P., Tipper, S.P.: Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin* 133(4), 694 (2007)
29. Gaver, W., Moran, T., MacLean, A., Lovstrand, L.: Realizing a Video Environment: EuroPARC's Rave System. In: *Proc. ACM CHI*, pp. 27–35 (1992)

30. Goffman, E.: *The presentation of self in everyday life*. Anchor Books, New York (1959)
31. Harrison, S.: *Media space 20+ years of mediated life*. Springer, London (2009)
32. Heath, C., Luff, P., Sellen, A.: Reconsidering the virtual workplace: flexible support for collaborative activity. In: Proc. ECSCW, pp. 83–99 (1995)
33. Iqbal, S.T., Bailey, B.P.: Understanding and developing models for detecting and differentiating breakpoints during interactive tasks. In: Proc. ACM CHI, pp. 697–706 (2007)
34. Isaacs, E., Tang, J., Morris, T.: Piazza: A desktop environment supporting impromptu and planned interactions. In: Proc. ACM CSCW, pp. 315–324 (1996)
35. Kastner, S., Ungerleider, L.G.: Mechanisms of Visual Attention in the Human Cortex. *Annual Review of Neuroscience* 23, 315–341 (2000)
36. Kendon, A.: *Conducting interaction: patterns of behavior in focused encounters*. Cambridge University Press, Cambridge (1990)
37. Kraut, R., Egido, C., Galegher, J.: Patterns of Contact and Communication in Scientific Research Collaboration. In: Proc. ACM CSCW, pp. 1–12 (1988)
38. McCrickard, D.S., Czerwinski, M., Bartram, L.: Introduction: design and evaluation of notification user interfaces. *Intl. Journal of Human-Computer Studies* 58, 509–514 (2003)
39. McEwan, G., Greenberg, S.: Supporting social worlds with the community bar. In Proc. ACM GROUP, pp. 21–30 (2005)
40. McFarlane, D.C., Latorella, K.A.: The scope and importance of human interruption in human-computer interaction design. *Human Computer Interaction* 17(1), 1–61 (2002)
41. Nummenmaa, L., Calder, A.J.: Neural mechanisms of social attention. *Trends in Cognitive Sciences* 13(3), 135–143 (2008)
42. Olguin, D.O., Gloor, P.A., Pentland, A.: Capturing individual and group behavior with wearable sensors. In: Proc. of the AAAI Spring Symposium on Human Behavior (2009)
43. Rodden, T.: Populating the application: a model of awareness for cooperative applications. In: Proc. ACM CSCW, pp. 87–96 (1996)
44. Schmidt, K.: The problem with 'awareness'. In: *Computer Supported Cooperative Work*, vol. 11, pp. 285–286 (2002)
45. Sudnow, D.: Temporal parameters of interpersonal observation. In: Sudnow, D. (ed.) *Studies in Social Interaction*. Free Press, New York (1972)
46. Tang, J.: Approaching and leave-Taking: Negotiating Contact in Computer-Mediated Communication. *ACM TOCHI* 14(1), 1–26 (2007)
47. Whittaker, S., Frohlich, D., Daly-Jones, O.: Informal Workplace Communication: What is It Like and How Might We Support It? In: Proc. ACM CHI, pp. 131–137 (1994)

Do Teams Achieve Usability Goals? Evaluating Goal Achievement with Usability Goals Setting Tool

Anirudha Joshi and N.L. Sarda

Indian Institute of Technology, Bombay
{anirudha,nls}@iitb.ac.in

Abstract. Do teams achieve important usability goals most of the time? Further, is goal achievement uniform or are practitioners more mindful of some goals than others? This paper presents an empirical study on usability goal achievement in industry projects. We used Usability Goal setting Tool (UGT), a recommender system that helps teams set, prioritize, and evaluate usability goals. The practitioner creates profiles for the product and its users. Based on these inputs, UGT helps the practitioner break down high-level usability goals into more specific goal parameters and provides recommendations, examples, and guidelines to assign weights to these parameters. UGT suggests strategies to evaluate goal parameters after the design is ready and assign them scores. UGT was used to collect data from 65 projects in the Indian software industry in which participants assigned weights and scores to the goal parameters. The 30 goal parameters suggested by UGT were found to be internally reliable, and having acceptable granularity and coverage. It was observed that goal parameter weights and scores correlated, but only moderately. Another interesting observation was that more than a third of the important goal parameters did not score well. We identify eight goal parameters that are typically high-weighted but have poor weight-score correlations. We call these “*latent but important*” goal parameters. Design teams will do well to pay closer attention to these goal parameters during projects.

Keywords: Usability goals achievement, usability goal parameters, latent goals, design tools, methods.

1 Introduction

Setting goals is an important step early in the design process. Setting goals before design gives the team a target to achieve. Goals help guide the design process, make the design activity tangible, and help evaluate the designs. In the field of human-computer interaction (HCI), often multi-disciplinary teams are involved; hence, setting goals early and getting an agreement from all stakeholders is important.

Goals have been discussed extensively in HCI literature [1-9]. To help teams set goals, we developed a Usability Goal Setting Tool (UGT) [10], [11]. UGT maintains a repository of profiles of past projects, their users, goals, and scores. Based on

this experience, UGT provides guidance to HCI practitioners to set and evaluate usability goals in new projects. Is such a tool necessary or is the current knowledge, skills and practices sufficient to let HCI practitioners set usability goals and evaluate products against them? During our interactions with software development teams in the industry, we observed that HCI practitioners are often unclear about the usability goals of the product they are designing. In some cases, the HCI practitioners are clear, but there is no explicit agreement on these by the other stakeholders in the team.

The questions we investigate in this paper are: To what extent do teams achieve their usability goals? Do teams achieve the important usability goals most of the time? Further, is goal achievement uniform or are practitioners more mindful of some usability goals than other?

We collected retrospective data with the help of UGT from projects in the Indian industry to evaluate goal achievement. We explored the data to establish the coverage, relevance, granularity, and internal reliability of the goal parameters.

In section 2, we review literature related to goal setting in design and in HCI literature. Section 3 gives an overview of UGT. In section 4, we present the data collected from industry projects and investigate the coverage, relevance, granularity, and internal validity of goals suggested in UGT. Section 5 investigates the goal achievement in these projects. Section 6 talks about conclusions and future work.

2 Goals in Design and HCI

The importance of goal-driven design has been stressed in literature for a long time. Design for a ‘need’ has been a part of traditional industrial design thinking. Charles Eames reportedly said, “*Design is a plan for arranging elements in such a way as best to accomplish a particular purpose*” [12]. Archer explains that since design is necessarily associated with change, identifying goals means “*defining the needs and pressures, which constitute the driving force for change*”. The first step is to determine the goals of the design effort together with “*the essential criteria by which a ‘good’ solution will be distinguished from a ‘not so good’ solution*” [13].

The closely related fields of HCI, usability, interaction design, and information architecture emphasise the importance of **usability goals**. ISO 9241 defines usability as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use [2]. Shneiderman and Nielsen agree on five high-level usability attributes of a product – learnability, speed of use, error-free use, retention over time, and subjective satisfaction [7], [1]. Mayhew categorizes usability goals as qualitative and quantitative, and as performance goals, preference goals, and satisfaction goals [3]. ISO 9126-1 describes usability in terms of understandability, learnability, operability, and attractiveness [4]. Cooper and Reimann emphasise the importance of goals in usability and interaction design process [6]. Hornbæk et al. suggest a technique to integrate business goals in usability evaluations [9]. Bevan summarises several other ways of organising usability measures, which could be looked upon as goals [8].

One purpose of defining these usability goal-sets was to reduce the ambiguity associated with the terms “*usability*” or “*user-friendliness*”, and define them in terms of measurable components. Usability has been recognised to be a multi-dimensional

property that needs to be broken down into measurable attributes such as learnability, speed of use etc. [1].

However, these attributes themselves are still at too broad to be interpreted consistently in the context of a given project. For example, one can interpret *learnability* as “users should take less time to learn to use the product” or as “users should be able to learn to use the product on their own”. *Error-free use* could be either interpreted as “product should not induce errors”, “product should tolerate user’s errors”, or “product should help the user recover from errors”. These interpretations could lead to communication gaps between team members and could make it difficult to evaluate the design. As summarised below, our approach in UGT is to split high-level usability goals into more granular, measurable, and less ambiguous goal parameters.

Relatively recent research interest in human emotions has broadened the traditional focus of researchers from usability to **user experience**. However, approaches to user experience goals are not universal. Unlike in usability, where high-level goals have been more or less agreed upon, when it comes to user experience design, there seems to be much less agreement. After a survey of user experience professionals, Law et al concluded that the concept of user experience is dynamic, context-dependent, and subjective [14]. There is no single accepted definition of what user experience is, and some researchers question whether the user experience could be designed at all [15].

While there may be no explicit agreement on a list of user experience goals, most practitioners favour a “goal-driven” approach in design of interactive products, e.g. [3], [5], [6]. When a practitioner sets goals in a particular project, he may not explicitly differentiate between whether a particular goal is related to usability of the product or its broader user experience.

In a typical industrial context, it may not be possible to meet all goals and it is necessary to **prioritize**. Depending on the context, users, and platform, some goals may be more important for a project, while others might be irrelevant. The need of prioritising goals has been acknowledged for a while. For example, Archer talks about “rank ordering” sub-problems as a method of prioritizing goals and resolving conflicts [13]. More recently, Cross talks about an objective tree method – organizing objectives into a hierarchy of higher and lower level objectives [16]. Shneiderman states, “a clever design for one community of users may be inappropriate for another community”, and “an efficient design for one class of tasks may be inefficient for another class” [7]. However, other than UGT (which we summarise below), there has been no guidance to prioritize usability goals systematically.

In our earlier work, we found that goal achievement is affected by the process. When HCI activities are better integrated with software development, teams following Agile processes achieve their usability goals better than teams following the Waterfall process [17]. We also found that unless they integrate HCI activities with software development rigorously, companies providing software development services do worse in usability goals achievement than companies involved in software product development. Elsewhere, we report the relative contributions of different HCI activities on goal achievement [18].

Despite the extensive literature on usability goals, there have been no empirical studies on the **extent of usability goals achievement**. We also do not know if some usability goals are achieved more frequently than others are. Further, it is not clear whether the current understanding of goals among practitioners is sufficient, or tools

such as the UGT are necessary to improve goal achievement. In this paper, we present one such study.

3 UGT Overview

The primary purpose of UGT is to help set usability goals, given a context of user goals and business goals [10]. This section provides an overview of UGT. A more detailed version of UGT including systematic instructions is available online [11].

UGT is envisioned as a recommender system (Fig. 1). Based on prior experience of similar product and user profiles, UGT recommends priorities for usability goals. UGT also recommends usability evaluation guidelines. At all times, the practitioner has the freedom to add goals, edit, or re-word the suggested goals, to regroup goals, or to evaluate them differently than suggested in UGT. These decisions are collected and fed back into UGT to improve its recommendations in future, making UGT flexible and extensible.

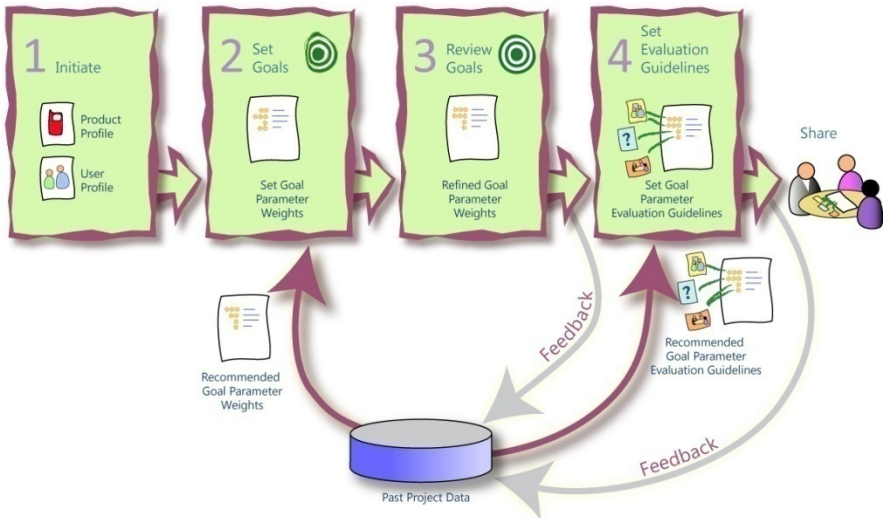


Fig. 1. An overview of UGT usage workflow

3.1 Goal Parameters in UGT

UGT helps break down high-level usability goals into 30 concrete, specific, and measurable goal parameters (Table 1). These were derived from the usability attributes suggested by Shneiderman [7] and Nielsen [1], and elaborated through brainstorming and formative evaluations as discussed in [10]. Our attempt is to keep these goal parameters granular, concrete, and yet relevant to a large number of projects.

The HCI practitioner determines the importance of each goal parameter in the context of his project by assigning it a suitable weight. Weights express the relative importance of goal parameters in the context of a project. For example, a product

meant to be used several times a day is likely to have higher weight for a speed related goal parameter such as goal parameter 8, “*user must be able to do primary task / frequent tasks quickly, easily, at all times*”. A one-time use product might give higher weight to learnability and ease of use goal parameters such as 1, “*findability: options / data / information should be visible / easy to find*” and 17, “*intuitiveness: user should be able to predict the next step / task*”. A life-critical product is likely to rate highly error-free use goal parameters such as 25, “*product should not induce errors*” and may sacrifice a learnability goal parameter such as 3, “*users should be able to learn on their own*”.

We suggest the following scale for assigning weights: Extremely important / unique selling proposition (5), very important (4), important (3), usual relevance / hygiene factor (2), somewhat relevant (1), irrelevant (0). Based on experience of prior projects, UGT recommends weights for the goal parameters on this scale [11].

We do not claim that the goal parameters and weights recommended by UGT cover all aspects of user experience in every situation. They are suggestive of the kind of goals that the practitioner *could* set. An experienced practitioner may assign different weights, add, edit, or re-word the suggested goal parameters, or group goal parameters differently if preferred. Yet, as we report below, the current list of goal parameters seems to have wide coverage, reasonable relevance, and sufficient granularity.

Goal-setters should be aware that while it may be tempting to set a high weight to each goal parameter, it might not be necessary, practical, or even possible to achieve such a design. The weights should reflect the priorities of the business, the stakeholders, and the users. The weights would help prioritize usability evaluation activity as the highest rated goals and parameters would be evaluated more rigorously.

After the evaluation, practitioners can assign a score to each goal parameter. We suggest the following scale for assigning scores: Outstanding, exceptional, best possible user experience against this goal parameter (100); acceptable, good enough, though not exceptional (75), undecided, neither good nor bad (50), bad, but not the worst (25), very bad, worst possible user experience for this goal parameter (0).

Table 1. Goals and goal parameters in UGT

Learnability	
1	Findability: options / data / information should be visible / easy to find
2	User should take less time to learn: (e.g. in < 10 minutes, in < 2 hours practice)
3	Users should be able to learn on their own
4	Product should be internally consistent
5	Product should be consistent with other products , older methods / past habits
6	Product should be consistent with earlier version
7	User should remember / retain critical, but infrequent tasks
Speed of use	
8	User must be able to do primary task / frequent tasks quickly, easily, at all times
9	User should be able to navigate quickly and easily
10	Product should not load user’s memory / product should not put cognitive load
11	Flexibility: User should control the sequence of tasks

Table 1. (Continued)

12	User should be able to complete tasks in specific time / no. of steps / in less efforts
13	Product should be personalised for the user automatically
14	Product should be localised for specific market segments
15	User should be able to customise the product for himself
Ease of use	
16	Interface should clearly communicate the conceptual model
17	Intuitiveness : User should be able to predict the next step / task
18	No entry barrier : user must be able to complete critical first tasks
19	Product should require no unnecessary tasks
20	Product should automate routine tasks / minimise user task load
21	Product should be always on , always accessible
Ease of Communication	
22	Information architecture : well aggregated, categorised, presented
23	Communication should be clear / user should easily understand text, visuals
Error-free use	
24	Product should give good feedback / display its current status
25	Product should not induce errors
26	Product should tolerate user's errors / forgiving interface / should prevent errors
27	Product should help user recover from errors / help users troubleshoot problems
Subjective Satisfaction	
28	User should feel in control of the product / behavioural appeal
29	User should feel emotionally engaged / brand / fun / reflective appeal / trust
30	User should find the product aesthetically appealing / visceral appeal
Other goals	
Add goal parameters specific to your project here.	

3.2 Using UGT

The HCI practitioner approaches UGT when he is clear about the product brief and has sufficient understanding about the domain, the problems, the context, and the users. Using UGT involves four steps: initiate, set goals, review goals, and set evaluation guidelines.

To begin with, UGT asks the practitioner to specify the **product profile** and one or more **user profiles**. The product profile captures information such as the industry domain, the work-practice domain, the expected cost to the user, the platforms, and the number of user profiles. Each user profile captures further information such as the age, the expected level of tech-savvyness of the user, the frequency of use, the product complexity, and the motivation of the user to use the product.

The product profile and the user profile form the input to UGT. Based on these, UGT provides recommendations for **assigning weights** to the goal parameters. Detailed recommendations for all the 30 goal parameters are available online [11].

These recommendations have been derived from experience of projects for which UGT has been used so far.

After the practitioner has done one pass of assigning weights to goal parameters, the list of goal parameters is re-presented, this time sorted by their weights. Goal parameters with weights that deviate substantially from the recommended range are highlighted. The practitioner **reviews** the weights and tweaks them. This step was introduced because it was observed during evaluations that practitioners over-assigned weights in the first pass, but preferred to tone them down during a review.

UGT recommends **evaluation guidelines** of each goal parameter. It recommends at least one possibility for a user-based test and one for a review-based evaluation. The evaluation guidelines for the 30 goal parameters are available online [11]. These are based on the formative evaluation reported in [10], and further data from projects where UGT has been applied so far.

As shown in Fig. 1, UGT has self-learning capabilities. The practitioner may use UGT recommendations or choose to override them. These decisions are captured and fed back to enrich UGT further. As UGT collects data from more projects, it gives more accurate and relevant recommendations.

4 Data and Analyses

Using UGT, we collected data about goal parameter weights and scores from industrial projects. First, we evaluate whether UGT goal parameters have sufficient coverage, relevance, and granularity, and whether the weights and scores used in UGT are internally reliable. Next, we determined the extent to which the HCI practitioners achieved goals of their projects, and whether practitioners are likely to be equally mindful of all goals parameters.

HCI practitioners from the Indian IT industry were trained on UGT. Then they were invited to participate in the study by contributing data from industry projects that they had recently completed. For each project, following data was collected:

- Product profiles and user profiles
- Goal parameter weights (0-5)
- Goal parameter scores (0-100). Participants were encouraged to use information from usability tests, reviews, and user feedback where available while giving their scores.

44 HCI practitioners participated and contributed data from 65 projects. Some participants contributed data from more than one project. The participants came from a wide variety of companies including four large (25,000+ employees) companies engaged in developing software on contract, a few relatively smaller companies, some multi-national companies with large product development centres in India, and smaller product development companies. The projects represented a wide variety as well. These included different platforms (desktop, web and mobiles), industry domains (finance, telecom, entertainment etc.), targeted users (call centre agents, sales persons, farmers etc.), types of business models (contracted software development companies and product companies), and process model used (waterfall and agile).

Table 2 lists the number of projects that assigned a particular weight (from 0 to 5) and a particular score (0-25-50-75-100) to each goal parameter. It also lists the means and standard deviations of weights and scores of each goal parameter.

Table 2. Number of projects that assigned a particular weight to a goal parameter, goal parameter weight means and standard deviations, number of projects that assigned a particular score to a goal parameter, and goal parameter score means and standard deviations. # refers to the goal parameter numbers in Table 1. (n = 65)

#	No. of projects out of 65 with weights						Weights		No. of projects out of 65 with scores					Scores	
	0	1	2	3	4	5	Mean	SD	0	25	50	75	100	Mean	SD
1	0	0	3	13	30	19	4.00	0.83	0	1	16	38	10	71.92	16.83
2	1	5	12	24	17	6	3.06	1.13	0	7	22	31	5	63.08	19.82
3	5	6	15	14	19	6	2.83	1.40	5	5	16	36	3	60.38	24.56
4	1	4	17	19	17	7	3.05	1.16	2	4	18	28	13	67.69	24.09
5	5	5	19	18	8	10	2.75	1.41	4	7	31	17	6	55.38	24.40
6	37	3	8	6	6	5	1.32	1.75	34	1	14	14	2	30.38	34.09
7	10	12	11	17	10	5	2.31	1.52	8	7	26	22	2	51.15	25.93
8	3	3	5	14	28	12	3.49	1.28	4	2	10	33	16	71.15	25.86
9	1	1	12	11	30	10	3.51	1.11	0	7	11	32	15	71.15	22.63
10	1	2	23	14	17	8	3.05	1.18	4	10	15	33	3	58.08	25.04
11	7	8	16	19	11	4	2.48	1.37	6	9	22	25	3	53.85	25.86
12	4	1	13	20	15	12	3.18	1.33	2	3	22	31	7	64.62	21.60
13	15	10	12	6	17	5	2.23	1.70	19	9	12	21	4	43.08	33.80
14	17	12	8	10	9	9	2.14	1.79	20	5	18	13	9	44.62	35.77
15	29	9	6	8	9	4	1.55	1.73	33	6	9	13	4	30.38	35.21
16	0	5	8	18	21	13	3.45	1.17	0	4	20	37	4	65.77	17.44
17	1	4	11	25	18	6	3.12	1.10	0	3	27	27	8	65.38	19.11
18	7	6	10	19	16	7	2.80	1.47	6	8	20	25	6	56.54	27.34
19	3	2	19	26	12	3	2.78	1.08	1	9	16	32	7	63.46	23.00
20	10	7	12	16	16	4	2.51	1.51	9	8	22	23	3	51.15	27.75
21	10	6	16	10	10	13	2.66	1.70	13	6	11	26	9	54.62	33.92
22	0	1	3	24	18	19	3.78	0.98	0	3	15	34	13	71.92	19.52
23	0	0	5	14	29	17	3.89	0.89	0	3	17	31	14	71.54	20.19
24	0	0	10	21	22	12	3.55	0.97	1	5	19	27	13	67.69	23.27
25	3	0	17	24	14	7	3.03	1.16	3	6	22	30	4	60.00	22.88
26	3	6	20	24	10	2	2.58	1.10	4	7	28	20	6	56.54	24.72
27	2	5	19	18	16	5	2.86	1.20	3	7	24	24	7	59.62	24.48
28	1	2	9	28	19	6	3.23	1.01	3	2	15	40	5	66.15	21.39
29	8	12	10	15	12	8	2.54	1.58	11	7	24	17	6	50.00	29.97
30	2	6	18	21	14	4	2.78	1.17	2	9	20	26	8	61.15	24.62

4.1 Goal Parameter Coverage, Relevance, Granularity and Reliability

While recommending goal parameters in UGT, our aim was to achieve a reasonable level of granularity so that the goal parameters are concrete, specific, measurable, and interpreted unambiguously. At the same time, we wanted the goal parameters to be widely applicable, relevant to a large number of projects, and internally reliable. We evaluated the data from projects to verify if this was indeed the case.

Coverage is the extent to which the goal parameters suggested by UGT suffice the goal-setting needs of a project. One method of evaluating if UGT has enough coverage was to look at how many new goal parameters were added by practitioners. Throughout the study, the participants had the freedom to add, edit, merge, or regroup goal parameters if the suggested ones did not suit their needs. This was indicated to them at the beginning of the study. To remind the participants, the UGT form always had blank lines under each group of goal parameters and at the bottom as shown in Table 1. After assigning goal weights, the participants were encouraged to express goals relevant to their projects beyond the ones listed in UGT.

A majority of the participants did not specify additional goal parameters. The goal parameters currently recommended in UGT could express goals in 59 out of 65 projects to the satisfaction of the practitioners in those projects. Practitioners of only 6 projects added goal parameters. Among these, 4 specified one additional goal parameter each, and 2 specified 2 additional goal parameters each. We can conclude that currently UGT has a reasonable coverage and is applicable to a majority of projects in the industry.

Relevance is the number of goal parameters that could be removed from UGT without hampering coverage significantly. Table 2 lists the number of projects that assigned a particular weight to each goal parameter. Out of the (65 projects x 30 goal parameters=) 1,950 cases, a weight greater than 0 is assigned to 1,757 (90%) cases and in only 193 (10%) of the 1,950 of the cases, the weight was set to 0 (irrelevant). As we can see in Table 2, weight 0 is not restricted to a few goal parameters, but spread across several. Further, even goal parameters weighted irrelevant by a large number of projects (goal parameter 6 and 15, for instance) have been weighted important (3 or more) by many other projects. We can conclude that the goal parameters suggested in UGT are relevant to projects.

Granularity is the extent to which UGT helps break down high-level goals into concrete goal parameters. Our aim was to achieve a good balance of granularity. We wanted the goal parameters to have sufficient granularity to allow practitioners to express themselves precisely, and yet not so fine that they cannot differentiate between adjacent elements. If the list of suggested goal parameters were too long, it could hamper the usability of UGT itself.

We could consider the granularity of UGT in two ways; the granularity of the goal parameters and the granularity of the scale (weights 0-5).

During the study, the participants seemed comfortable with the scale. Assigning a specific meaning to each point on the scale seems to have helped. Once they got used to it, the participants could easily differentiate between the adjacent points on the scale and could justify why one goal parameter would have a weight of 2 while another would have a weight of 3.

Granularity of the goal parameters is harder to evaluate. In some cases, a few of the goal parameters could be split up further and made more granular, e.g. goal parameter 29 “*user should feel emotionally engaged / brand / fun / reflective appeal / trust*”. On the other hand, since none of the participants felt the need for doing so for their projects though they had a choice, it is perhaps not necessary.

Our objective behind splitting high-level goals into goal parameters is to make goal setting less ambiguous and consistently interpretable. During the study, participants could interpret the goal parameters unambiguously. Occasionally, they needed to refer to the UGT documentation and go through examples for some of the goal parameters. Nevertheless, once they understood the meaning, they could unambiguously interpret the goal parameter in the context of their project, determine its importance, and think of ways to evaluate that goal parameter. We believe that we have achieved a right balance of granularity in UGT goal parameters. Of course, in case a particular project needs more granular goal parameters, UGT always allows for the flexibility.

When variables are used as components of a larger construct, their internal consistency reliability is important. **Internal consistency reliability** is the extent to which individual variables are measuring the same underlying construct of interest [19]. Cronbach’s alpha evaluates the variation accounted for by the true score of the underlying construct. Cronbach’s alpha above 0.7 is considered an acceptable measure of internal reliability. To test if any particular component varies differently than the other components, and therefore does not measure the same underlying construct, that component is dropped and alpha is calculated again with remaining components. If the resulting alpha increases substantially, then that component is not measuring the same underlying construct and it must be deleted.

The overall internal reliability of the 30 goal parameters in the 65 projects using Cronbach’s alpha was 0.7870 for the weights, and 0.8733 for the scores. Both values indicate an acceptable level of internal reliability. Each of the 30 goal parameters was deleted by turns and Cronbach’s alpha was re-calculated for the remaining 29 goal parameters. The resulting alphas did not vary much – from 0.7720 to 0.7969 for weights and from 0.8629 to 0.8780 for scores. The alpha resulting from dropping any particular goal parameter did not increase by much (maximum increase < 0.01). We can conclude that the 30 goal parameters are internally consistent and essential to measure the same construct.

5 Goal Achievement

We analysed extent to which the HCI practitioners achieve goals of their projects. In an ideal situation, design teams would like to score well against all goal parameters. In practice, if resources were limited, design teams would still like to score well on high-weighted goal parameters. This analysis was done to test the following hypothesis: **Design teams naturally concentrate on achieving the high-priority goals; goal weights and goal scores are positively correlated.**

This hypothesis can be tested in two different ways. We can evaluate the correlation between the mean weights and mean scores of the 30 goal parameters reported in Table 2. However, mean weights and scores can disguise the divergence between

individual weight-score pairs. Therefore, we can also evaluate the correlations between individual weight-score pairs (65 projects x 30 goal parameters = 1,950 pairs).

Pearson's correlation r is used to evaluate correlations between parametric variables, where r^2 indicates the extent of variation in the dependent variable because of the independent variable [20]. Pearson's correlation of mean weights and mean scores of the 30 goal parameters (Table 2, $n = 30$) was calculated. A significant positive and strong correlation resulted ($r = 0.957$, $n = 30$, $p < 0.0005$, $r^2 = 0.916$). We can conclude that the mean weights of goal parameters strongly correlate with mean scores of goal parameters, and mean weight of goal parameters predicts 91.6% variation in mean scores.

Next, a crosstabs summary was generated for the 1,950 weight-score pairs (Table 3). This shows how many goal parameters achieved each weight-score pair. For example, among the goal parameters that were set a weight of 5, 52 achieved a score of 100, while 97 achieved a score of 75. We can see that out of 1,950 goal parameters 757 (43%) scored 50 or less (undecided or worse). A matter of particular concern is that goal parameters with weights 3, 4 and 5 (the important goal parameters) scored 50 or less in 39%, 36% and 33% cases respectively (the gray cells in Table 3).

Table 3. SPSS crosstabs output for 30 goal parameter weights and scores for 65 projects. The highlighted cells (important, but low-scoring goal parameters) are a matter of concern.

		Goal score					Total
		0	25	50	75	100	
Goal weight	0	144 (75%)	9 (5%)	19 (10%)	19 (10%)	2 (1%)	193 (100%)
	1	21 (14%)	35 (24%)	44 (30%)	37 (25%)	10 (7%)	147 (100%)
	2	17 (5%)	41 (11%)	142 (39%)	146 (40%)	23 (6%)	369 (100%)
	3	10 (2%)	37 (7%)	157 (30%)	254 (49%)	60 (12%)	518 (100%)
	4	11 (2%)	28 (6%)	138 (28%)	256 (51%)	65 (13%)	498 (100%)
	5	1 (0%)	18 (8%)	57 (25%)	97 (43%)	52 (23%)	225 (100%)
Total		204 (11%)	168 (9%)	557 (29%)	809 (42%)	212 (11%)	1,950 (100%)

Spearman's rho is used to evaluate correlations between ordinal variables and the square of the Spearman's rho indicates the extent of variation in the dependent variable because of the independent variable [20]. As the goal parameter weights and scores are ordinal variables, a Spearman's rho was calculated between the 1,950 pairs of weights and goal parameter scores reports. A statistically significant correlation emerged but the value of the rho was quite small ($\rho = 0.391$, $n = 1,950$, $p < 0.0005$).

A correlation value between 0.25 and 0.5 is considered a moderate association. The $\rho^2 = 0.153$. This indicates that only 15.3% of the variation in the goal parameter scores is explained by the variation in the goal parameter weights.

While mean weights and mean scores of goal parameters correlate very well, individual goal parameter weight-score pairs correlate only moderately, and as many as 36.8% of the important goals are not achieved. This could have happened possibly because the practitioners did not set usability goals systematically during the project, or because they did not track goal achievement of important goals explicitly at the end of the project. As a result, they may have not explicitly realised the importance of some of the goal parameters for their projects until they participated in our study.

A limitation of this study was that data was collected retrospectively, i.e. after the projects were completed, and this may have introduced some biases in the data. However, this may not be significantly affecting our main finding that a large proportion of important usability goals are not achieved. It is anecdotally observed that practitioners set more ambitious goals in the beginning of the project than at the end. If at all, it could be possible that in reality an even larger proportion of important usability goals are not achieved.

5.1 Latent Goals and Explicit Goals

If the practitioner knew that a goal parameter is important in the context of a project, he would strive to achieve it well. In our study, he would assign that goal parameter a high weight. If he were successful in achieving that goal parameter, he would also assign it a high score.

A high correlation between the weights and scores of a particular goal parameter across projects (Fig. 2a) indicates that largely, practitioners put in efforts to achieve high scores against that goal parameter in projects where it was important. Conversely, a low correlation (Fig. 2b) indicates that scores of that goal parameter are independent of its weights. This could happen if either the goal parameter is either particularly easy or difficult to achieve (and hence uniformly high or low scores), or if the practitioners were not mindful of that goal parameter during projects.

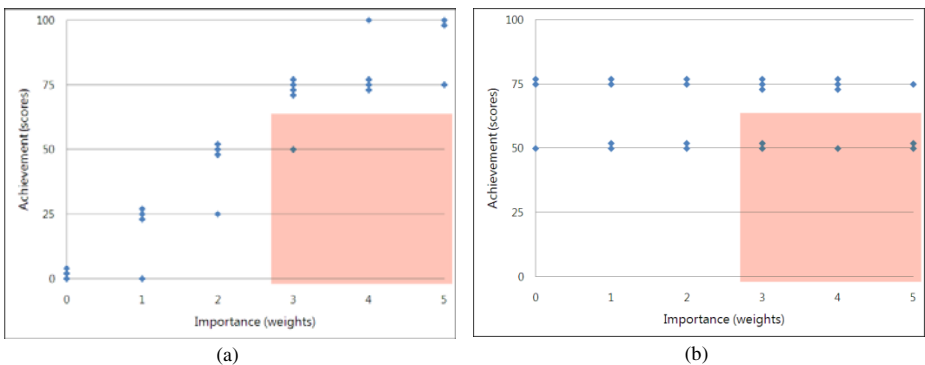


Fig. 2. Importance / achievement scatter plot for (a) a highly correlated goal parameter, and (b) a non-correlated goal parameter (hypothetical data). Each dot represents data from one project. The pink area represents the projects that gave weight of 3 or more, but scored 50 or less.

If practitioners were equally mindful of all goal parameters during development, there would not be a large variation between weight-score correlations across goal parameters. On the other hand, if some goals tended to be “*latent*” (i.e. the practitioners were less mindful of them during development), these would tend to have lower weight-score correlations. This analysis was done to test the following hypothesis: **Design teams are equally mindful of all goal parameters; weight-score correlations do not vary much across goal parameters.**

Spearman’s rhos were calculated to explore the relationship between weights and scores for individual goal parameter across the 65 projects. Table 4 lists the rhos and significances for the 30 goal parameters, along with mean weights and mean scores. The statistically significant rhos ($p < 0.05$) are highlighted in grey. The table has been sorted in the descending order of mean weights assigned to each goal parameter.

Table 4. Spearman’s rhos for weights and scores, rho significances of individual goal parameters (n = 65 projects each). The mean weights and mean scores are reproduced here from Table 2 for convenience.

Goal parameters (see Table 1 for longer names)	Spearman’s rho for weight – score	Significance of rho $p \leq$	Mean weights (0-5)	Mean scores (0-100)
1 Findability	0.17	0.186	4.00	71.92
23 Clear communication	-0.19	0.138	3.89	71.54
22 Info architecture	-0.12	0.335	3.78	71.92
24 Feedback	-0.02	0.901	3.55	67.69
9 Quick and easy navigation	0.09	0.482	3.51	71.15
8 Do primary tasks quickly	0.34	0.006	3.49	71.15
16 Conceptual model	0.03	0.834	3.45	65.77
28 Users feel in control	0.17	0.181	3.23	66.15
12 Complete tasks in time	0.17	0.186	3.18	64.62
17 Intuitiveness	0.23	0.070	3.12	65.38
2 Less time to learn	0.22	0.074	3.06	63.08
4 Consistent: internally	0.39	0.001	3.05	67.69
10 Memory / cognitive load	0.02	0.897	3.05	58.08
25 Not induce errors	-0.04	0.725	3.03	60.00
27 Help error recovery	0.21	0.092	2.86	59.62
3 Learn on their own	0.54	0.0005	2.83	60.38
18 No entry barrier	0.28	0.021	2.80	56.54
19 No unnecessary tasks	0.07	0.558	2.78	63.46
30 Aesthetic appeal	0.33	0.007	2.78	61.15
5 Consistent: other prods.	0.38	0.002	2.75	55.38
21 Always on	0.55	0.0005	2.66	54.62
26 Forgiving interface	0.35	0.004	2.58	56.54
29 Emotional engagement	0.53	0.0005	2.54	50.00
20 Minimise task load	0.48	0.0005	2.51	51.15

Table 4. (Continued)

11	Flexibility / user control	0.33	0.007	2.48	53.85
7	Retain infrequent tasks	0.24	0.059	2.31	51.15
13	Auto-personalised	0.71	0.0005	2.23	43.08
14	Localised	0.64	0.0005	2.14	44.62
15	User can customise	0.74	0.0005	1.55	30.38
6	Consistent: earlier version	0.77	0.0005	1.32	30.38

As reported (Table 2), the mean scores of goal parameters are between 30 and 72 and standard deviations of scores are greater than 16. This indicates that none of the goal parameters was particularly easy or particularly difficult to achieve. Yet, the weight-score rhos are not significant for half the goal parameters (Table 4) in spite of having a reasonable sample size of projects ($n = 65$). We can conclude that practitioners were not mindful of these goal parameters during projects, but gave them a higher weight during our study.

Interestingly, goal parameters with high mean weights tend to have less significant weight-score rhos (the upper half of Table 4). Conversely, the goal parameters with low mean weights tend to have higher and statistically significant weight-score rhos (the lower half of Table 4). The Pearson's correlation between the goal parameter weight means and the weight-score Spearman's rhos is strongly negative and statistically significant ($r = -0.815$, $n = 30$, $p < 0.0005$).

We can conclude that weights and scores are better correlated for low-weighted goal parameters, but not so well correlated on higher weighted goal parameters. Design teams achieve better scores when a typically low-weighted goal parameter gets an occasional higher weight, but do not achieve similar better scores when a typically high-weighted goal parameter gets an even higher weight.

A possible interpretation is that the typically high-weighted goals are **latent but important**. These goals are perhaps not explicitly asked for by the stakeholders and hence are perhaps not explicitly evaluated for during usability evaluations. The practitioners gave higher weights to these goal parameters when they saw them listed in UGT during the study, but they were not very mindful of them during the projects.

The following goal parameters have particularly low weight-score Spearman's correlations ($\rho < 0.15$) and are potentially latent but important. Two of these have been plotted in Fig. 3 as an illustration. These 8 goal parameters have a mean weight of 3.40.:

- 23 - Communication should be clear: $\rho = -0.19$ ($p = 0.138$) (Fig. 3a)
- 22 - Information architecture should be well aggregated, categorised, presented: $\rho = -0.12$ ($p = 0.335$)
- 25 - Product should not induce errors: $\rho = -0.04$ ($p = 0.725$)
- 24 - Product should give good feedback / display its current status: $\rho = -0.02$ ($p = 0.901$) (Fig. 3b)
- 10 - Product should not load user's memory / put cognitive load: $\rho = 0.02$ ($p = 0.897$)
- 16 - Interface should clearly communicate the conceptual model: $\rho = 0.03$ ($p = 0.834$)

- 19 - Product should require no unnecessary tasks: $\rho = 0.07$ ($p = 0.558$)
- 9 - Users should be able to navigate quickly and easily: $\rho = 0.09$ ($p = 0.482$)

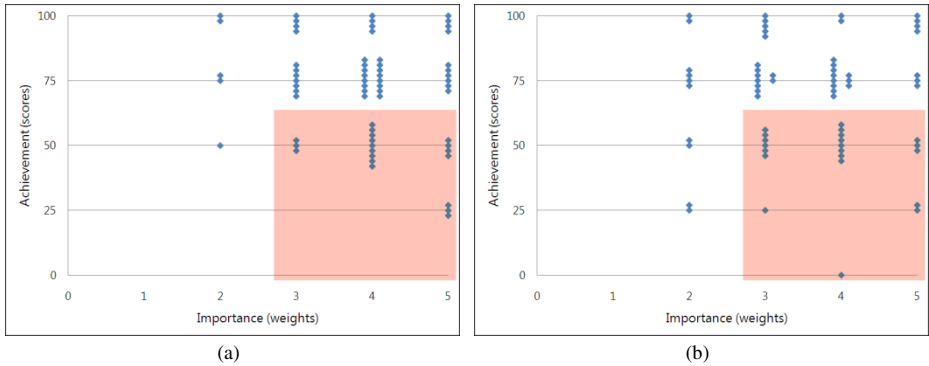


Fig. 3. Importance / achievement scatter plot for (a) Goal parameter 23 “communication should be clear”, mean weight = 3.89 (SD = 0.89), mean score = 71.54 (SD = 20.19), $\rho = -0.19$ ($p = 0.138$), and (b) Goal parameter 24 “product should give good feedback”, mean weight = 3.55 (SD = 0.97), mean score = 67.69 (SD = 23.27), $\rho = -0.02$ ($p = 0.901$) ($n = 65$ in each case). Each dot represents data from one project. The pink area represents the projects that gave weight of 3 or more, but scored 50 or less.

A corollary to the above would be that low-weighted, high correlation goal parameters are **explicit**. These are more readily expressed by the stakeholders (such as clients or product managers) and are specifically addressed during usability evaluations. The following goal parameters have high and significant weight-score Spearman’s correlations ($\rho > 0.50$, $p < 0.05$) and are potentially explicit. These 7 goal parameters have a mean weight of 2.21, which is significantly less than the weights of latent goals above ($p < 0.0005$). Two of these have been plotted in Fig. 4 as an illustration:

- 6 - Product should be consistent with earlier version: $\rho = 0.77$ ($p = 0.0005$) (Fig. 4a)
- 15 - User should be able to customise the product for himself: $\rho = 0.74$ ($p = 0.0005$)
- 13 - Product should be personalised for the user automatically: $\rho = 0.71$ ($p = 0.0005$) (Fig. 4b)
- 14 - Product should be localised for specific market segment: $\rho = 0.64$ ($p = 0.0005$)
- 21 - Product should be always on, always accessible: $\rho = 0.55$ ($p = 0.0005$)
- 3 - Users should be able to learn on their own: $\rho = 0.54$ ($p = 0.0005$)
- 29 - User should feel emotionally engaged with product / brand / product should be fun / reflective appeal / trust: $\rho = 0.53$ ($p = 0.0005$)

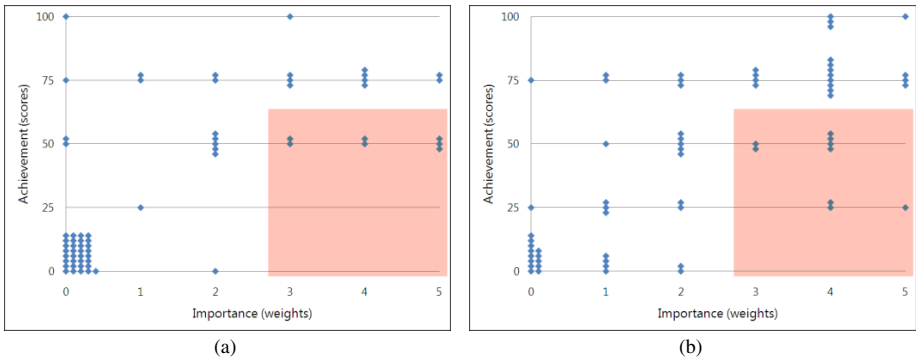


Fig. 4. Fig: Importance / achievement scatter plot for (a) Goal parameter 6 “consistent with earlier version”, mean weight = 1.32 (SD = 1.75), mean score = 30.38 (SD = 34.09), $\rho = 0.77$ ($p = 0.0005$), and (b) Goal parameter 13 “product should be personalised for the user”, mean weight = 2.23 (SD = 1.70), mean score = 43.08 (SD = 33.80), $\rho = 0.71$ ($p = 0.0005$), ($n = 65$ in each case). Each dot represents data from one project. The pink area represents the projects that gave weight of 3 or more, but scored 50 or less.

6 Conclusions and Future Work

We knew that there are differences in goal achievement in projects following waterfall and agile process models and between product and services companies [17]. We also knew that some HCI activities contribute to goal achievement more than others [18]. In this paper, we presented an empirical study that looks at goal achievement in 65 industry projects with the help of a tool for usability goal setting (UGT). In these projects, the goal parameter weights and scores correlated, though only moderately. HCI practitioners seem to be achieving important usability goals only moderately better than other usability goals. A particular matter of concern is that more than a third of the important usability goals scored undecided or worse. This clearly establishes the need to pay more attention to explicitly setting usability goals early on in the project, tracking the achievement of these goals during the project, and evaluating against these goals at the end.

We also identified the differences in goal achievement patterns between usability goals. Projects seem to score well on a typically less important usability goal when it occasionally becomes more important, but do not achieve better scores when a typically important usability goal becomes even more important. We identified these usability goals that could be interpreted as “important but latent”. Practitioners gave a higher weight to these goals when they saw them during the study, but judging by the scores of these goal parameters, we can only conclude that they were not very mindful of these higher weights during the projects.

Further research is needed in goal setting, tracking, and achievement to determine the extent to which goal achievement varies based on variables such as organisational maturity, domain, platform, and experience of practitioners.

The study established the usefulness of a tool such as UGT. In an earlier survey, we had reported that practitioners thought that UGT was useful, it helped

them understand the context of the project better, and made them think about goals that they had not considered earlier [10]. This study independently corroborates this result. It points to a need for a tool such as UGT that will help practitioners set and track goals during their projects systematically and provides guidance in prioritisation and evaluation of goals. As a long-term activity, organizations can use such a tool to keep a history of past goals, usability evaluations, practices, and costs, and use the repository as a knowledge base for identification and prioritization of usability goals in future projects. More studies will be required to determine if tools such as UGT could help improve usability goal achievement in practice. We also plan to explore the relationships between the user profiles, the product profiles, and the goal parameter weights with the aim of improving the recommendations for the goal parameter weights.

The study had some limitations that we acknowledge. The study comprised of projects from the Indian IT industry only. This could have introduced certain biases. A majority of the projects used the waterfall process model of software development, and a few used agile process models. No projects reported the use of Rational Unified Process models. This proportion may be different in other countries / contexts. The maturity of the usability, the mix of business models (product vs. service) and nature of domains addressed may also be different in other countries. Further, as discussed above, data was collected retrospectively. Thirdly, for most projects, only one representative participated in the study. This happened due to practical constraints. In reality, the data could vary somewhat if we get inputs from multiple stakeholders in each project as UGT is meant to take.

Acknowledgements. We thank all the participants who provided data and helped us evaluate the UGT. We thank Dr. Sanjay Tripathi and Prof. SV Sabnis for their help and feedback in the statistical analysis. We thank Prof. UA Athavankar, Prof. Umesh Bellur, and Prof. S Sudarshan of IIT Bombay for their suggestions in this project.

References

1. Nielsen, J.: Usability Engineering, pp. 26–37. Morgan Kaufmann, San Francisco (1993)
2. International Organization for Standardization: ISO 9241-1:1997 Ergonomic Requirements for Office Work with Visual Display Terminals (1997)
3. Mayhew, D.: Usability Engineering Lifecycle, pp. 123–145. Morgan Kaufmann, San Francisco (1999)
4. International Organization for Standardization: ISO/IEC 9126-1:2001 Software Engineering - Product Quality (2001)
5. Preece, J., Rogers, Y., Sharp, H.: Interaction Design, Beyond Human-Computer Interaction, 1st edn., pp. 13–20. John Wiley & Sons, Chichester (2002)
6. Cooper, A., Reimann, R.: About Face 2.0, pp. 64–67. Wiley, Chichester (2003)
7. Shneiderman, B.: Designing the User Interface: Strategies for Effective Human-Computer Interaction, 4th edn., p. 14. Addison Wesley, Reading (2004)
8. Bevan, N.: Classifying and Selecting UX and Usability Measures. In: International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (2008)
9. Hornbæk, K., Frøkjær, E.: Making Use of Business Goals in Usability Evaluation: An Experiment with Novice Evaluators. In: CHI, Florence (2008)

10. Joshi, A.: Usability Goals Setting Tool. In: 4th Workshop on Software and Usability Engineering Cross-Pollination: Usability Evaluation of Advanced Interfaces, Uppsala (2009)
11. Joshi, A.: Usability Goals Setting Tool - Online Version,
<http://www.idc.iitb.ac.in/~anirudha/ugt.htm> (accessed 2009)
12. Eames, C.: Charles Eames Quotes. In: BrainyQuote.com,
<http://www.brainyquote.com/quotes/quotes/c/charleseam169187.html>
13. Archer, B.: Systematic Method for Designers. Council of Industrial Design (1965)
14. Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., Kort, J.: Understanding, Scoping and Defining User eXperience: A Survey Approach. In: CHI 2009 (2009)
15. McCarthy, J., Wright, P.: Technology as Experience, pp. 49–129. The MIT Press, Cambridge (2004)
16. Cross, N.: Engineering Design Methods, 3rd edn., pp. 61–76. John Wiley & Sons, Chichester (2000)
17. Joshi, A., Sarda, N., Tripathi, S.: Measuring Effectiveness of HCI Integration in Software Development Processes. *Journal of Software Systems* (2010), doi:10.1016/j.jss.2010.03.078
18. Joshi, A., Sarda, N.: Evaluating Relative Contributions of Various HCI Activities to Usability. In: *Human-Centred Software Engineering*, Reykjavik (2010)
19. Hatcher, L.: A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling, pp. 131–140. SAS Publishing (1994)
20. Kurtz, N.: Introduction to Social Statistics, pp. 270–275. McGraw Hill Book Company, New York (1983)

Supporting Window Switching with Spatially Consistent Thumbnail Zones: Design and Evaluation

Susanne Tak¹, Joey Scarr¹, Carl Gutwin², and Andy Cockburn¹

¹ Computer Science and Software Engineering, University of Canterbury,
Private Bag 4800, Christchurch 8140, New Zealand
susanne.tak@pg.canterbury.ac.nz,
jls129@student.canterbury.ac.nz, andy@cosc.canterbury.ac.nz

² Department of Computer Science, University of Saskatchewan,
110 Science Place, Saskatoon, Saskatchewan, S7N 5C9, Canada
gutwin@cs.usask.ca

Abstract. Computer users switch between applications and windows all day, but finding the target window can be difficult, particularly when the total number of windows is high. We describe the design and evaluation of a new window switcher called SCOTZ (for Spatially Consistent Thumbnail Zones). SCOTZ is a window switching interface which shows all windows grouped by application and allocates more space to the most frequently revisited applications. The two key design principles of SCOTZ are (1) predictability of window locations, and (2) improved accessibility of recently and frequently used windows. We describe the design and features of SCOTZ, and present the findings from qualitative and empirical studies which demonstrate that SCOTZ yields performance and preference benefits over existing window switching tools.

Keywords: window switching, revisitation, spatial stability, predictability.

1 Introduction

Desktop computing involves constant window switching to navigate between various applications and documents. Previous work has found that people have more than eight windows open almost 80% of the time and that the average time between window switches is only 20.9 seconds [1].

Several window switching tools are available in most operating systems, and there are a large number of commercial and research tools that aim to enhance window switching performance. Our previous work identified two problems with current window switching interfaces [2]: first, many window switching interfaces lack spatial stability, meaning that the location of controls for acquiring particular windows can change over time, forcing users to rely on relatively slow visual search rather than rapid spatial decisions; second, most common window switching interfaces provide relatively weak support for strongly-exhibited patterns of window revisitation. We proposed spatial constancy and size morphing as design solutions to these problems, and our experiments using synthetic target acquisition tasks provided preliminary evidence of their success [2].

In this paper we extend our prior findings with the design and evaluation of a new window switcher called SCOTZ (Spatially Consistent Thumbnail Zones). SCOTZ uses application zones arranged in a treemap visualization to provide a spatially-stable and predictable window layout, and uses size morphing of the application zones to facilitate revisitation. It also provides quick access to recently used windows. SCOTZ's design is validated through qualitative and quantitative evaluations.

We make three specific contributions:

- a thorough description of our new window switcher SCOTZ and the rationale for its design choices;
- a summary of lessons learned from our qualitative study, which is useful for future iterations of SCOTZ as well being potentially informative for the design of other window switchers;
- an empirical study demonstrating the performance benefits of SCOTZ over two major commercial window switching interfaces (the Windows 7 Taskbar and Windows 7 Alt+Tab).

2 Related Work

2.1 Commercial Window Switching Interfaces

Three important commercial window switching applications are briefly reviewed below: the Microsoft Windows 7 Taskbar, Alt+Tab, and Mac OS X Exposé.

The *Microsoft Windows 7 Taskbar* is a narrow strip at the bottom of the screen which groups windows by application (see Fig. 1). Taskbar application icons can be *pinned* to the Taskbar so they are always in the same location and remain visible on the Taskbar even when there are no associated windows open. These pinned application controls remain in stable locations across sessions. However, the Taskbar provides no explicit support for switching to recent or frequently-used windows.

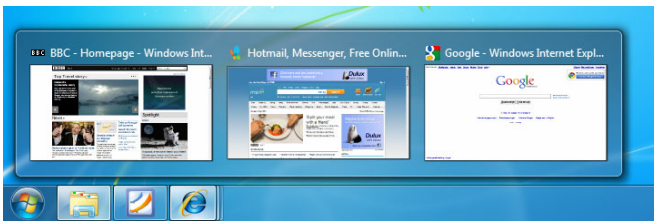


Fig. 1. The Microsoft Windows 7 Taskbar

Microsoft Windows 7 Alt+Tab (sometimes referred to as *Windows Flip*) is a modal window that is shown when the key combination Alt+Tab is pressed (see Fig. 2). The Windows 7 Alt+Tab window shows thumbnail previews of all windows: first, the top six windows in the z-order (see below), then a desktop preview, and then all remaining windows sorted by application name. Z-ordering is the depth-ordering of windows: a window that is placed in front of another window is relatively higher in the

z -ordering than the underlying windows. Z -ordering is similar to recency ordering, but sorting windows by z -order is spatially unstable: the ordering of the window representations will be different from switch to switch if the z -ordering of the windows changes. Also, it is unclear how well users understand and can anticipate z -order.



Fig. 2. Microsoft Windows 7 Alt+Tab

When activated, *Mac OS X Exposé* smoothly shrinks all windows so that they can be simultaneously viewed on the screen. The spatial location of each window in the overview is influenced by its most recent location on the screen. While this relative positioning may assist users in visually seeking windows in the overview, the locations are not spatially stable between invocations if the locations of the windows change, and consequently the locations are unpredictable.

2.2 Task-Based Window Switching Interfaces

Early studies of everyday computing [3] observed that computer users frequently switch between *tasks* such as writing a paper or programming. Many window switching interfaces have been developed to support *task management* by grouping windows by task. For example, a “writing a paper” task might contain a spreadsheet window, a statistical analysis package window, and a word-processing document into which the results are typed.

The Rooms system [4] is an early example of a window desktop management system that allows users to manually partition space for different tasks. Examples of task-based window managers that require *manual* grouping of windows into tasks are GroupBar [5], Activity-Based Computing [6], Scalable Fabric [7] and Task Gallery [8]. GroupBar and Activity-Based Computing use interfaces similar to the Windows Taskbar. With Scalable Fabric users can place groups of related windows in the periphery of the screen and switch to these groups of windows at once. Task Gallery uses a 3D visualization of task groups. The primary limitation of manual creation of task groups is that users must carry out explicit actions to gain potential benefits and it creates additional cognitive overhead for the user [9, 10].

Some window switching interfaces automatically identify tasks. SWISH [11] is an algorithm to automatically group windows into tasks based on their temporal and semantic relatedness. RelAltTab [12] modifies the Alt-Tab list to incorporate a system-generated list of related windows, similar to SWISH. Push-and-pull switching [13] automatically identifies window groups based on window overlap. The primary limitations of automatically-adaptive systems are that they can incorrectly predict the user’s intention and that users can fail to understand or anticipate the system’s adaptation [14]. When this happens users must resort to time-consuming visual search of candidate targets.

WindowScape [15] uses a combination of automatic grouping by taking snapshots of the desktop (which can then later be revisited) as well as manual configuration of window miniatures on the desktop.

Another issue related to task-based grouping of windows is whether or not windows can be associated with multiple tasks at the same time. Some windows may not be ‘task-specific’ at all. For example, generic applications, such as a web browser or an e-mail client are likely to be used across tasks, rather than in one specific task. If a task-based window switcher only allows for exclusive grouping (i.e., a window can only be associated with one task, e.g. [5, 6, 8]) the user is forced to make ‘impossible’ choices about where a window should belong, or open multiple windows for these applications, which some users find unnatural or difficult (as one user stated, “[I’m] still trying to get used to having multiple internet windows open” [5], p. 40). Using Activity-Based Computing [6], windows can be classified under more than one task, although the authors report on problems with achieving this in “a simple manner.”

Last, another problem associated with task-based grouping of windows is that some windows might not be associated with any task at all. This is reported in [6]: *“The worst thing? Well [...] if you have to put everything into activities, then you need to constantly consider ‘where does this one belong’. In many situations something just appears quickly and then you start up some application and do some things in it. [...]”* (p. 219).

Taskposé [16] uses a fuzzy approach to defining tasks, using spatial proximity to illuminate task-based window relationships – windows that are often temporally adjacent drift closer to one another and those that are temporally distant drift apart.

2.3 Studies of Window Switching

In this section, we provide an overview of studies related to window switching behaviour and window switching tools.

Window Switching Behaviour. Very early studies [17] of window switching showed that window switching is much more common than window creation, deletion and geometry management such as moving and resizing. Recent studies have showed the same: Hutchings et al. [18] found that the average time a window is active is only 20.9 seconds. In terms of *how* people switch between windows, previous work has found a relationship between monitor setup and window switching methods, with dual monitor users more likely to use a direct click on a window and less likely to use the Windows Taskbar, than single monitor users [2, 18]. In terms of *which* windows people switch to, previous work [2] has found that window switching follows an inverse exponential distribution, with 80% of window switches involving only 35% of windows.

Window Switching Interfaces. Analysis of the efficiency and effectiveness of current commercial methods for switching between windows is often anecdotal and sometimes conflicting. For example, some previous work labels the ordering of windows in Alt+Tab as “very effective” [19], but Alt+Tabbing is also labelled as “tedious” [20]. There are very few studies that evaluate the use, efficiency and/or effectiveness of the common commercially available window switching interfaces in

a formal manner. Alt+Tab is found to be relatively fast when the number of windows is small [21], but performance decreases as the number of windows increases. Users make more errors when using the Windows Taskbar than when using Mac's Exposé, which may be due to the smaller target sizes on the Windows Taskbar [21].

3 SCOTZ

In this section, we introduce our new window switcher SCOTZ (Spatially Consistent Thumbnail Zones)¹. SCOTZ is designed to satisfy two main goals:

1. **Predictability of window locations.** Comprehensibility and predictability are important attributes of a user interface [14]. If a window switcher places window representations in predictable locations the user will be able to correctly anticipate where a window representation is (or will be) placed, reducing the need for a costly visual search.
2. **Improved accessibility of frequently/recently used applications and windows.** Previous work [2] has identified strong revisitation patterns to applications and windows, so window switching interfaces should support rapid acquisition of recent or frequent targets.

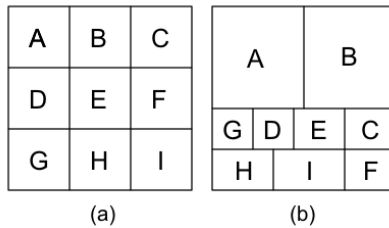


Fig. 3. Abstraction of the design of SCOTZ with nine application zones, labelled A to I: (a) before size morphing, (b) after several iterations of size morphing; applications A and B have been switched to frequently, and are therefore allocated more space, while keeping all application zones in relatively stable positions.

SCOTZ is a modal interface which groups windows by application in so-called *application zones*. Size morphing is applied to allocate more space to the most frequently switched-to applications. The basic concept behind SCOTZ is shown in Fig. 3, and Fig. 4 shows an actual SCOTZ window in full-screen mode with eight application zones and six windows. SCOTZ retains application zones' size and position even when the computer has been restarted.

The following sections describe the design of SCOTZ and explain how its features fulfil the two main goals above. Also, we describe four additional properties of the system: support for different display sizes, support for keyboard and mouse input, support for application launching, and options for personalization.

¹ A fully functional version of SCOTZ is available for download at <http://www.cosc.canterbury.ac.nz/scotz>

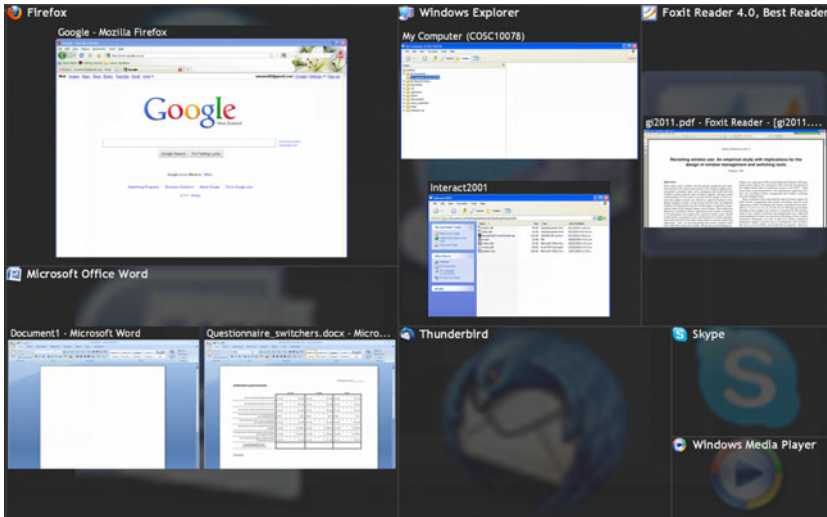


Fig. 4. SCOTZ in full-screen mode showing eight application zones and six windows

3.1 Predictable Window Locations

SCOTZ groups windows by application in *application zones*, which remain in stable locations. This results in comprehensible and predictable window locations. SCOTZ offers three different layouts for the application zones: two treemap layouts, where size morphing is applied, and a grid layout where no size morphing is allowed (see Fig. 5). A treemap is a space-filling layout that recursively divides the screen in rectangles, sized relative to some underlying data attribute [22]. In SCOTZ, the size of the application zone reflects how often an application has been switched to.

Various algorithms for generating treemaps exist, each offering advantages for specific contexts. For example, some treemaps are designed to keep items in relatively stable positions as the underlying data changes. An example of such a treemap is the spiral treemap [23] (see Fig. 5a), which preserves the ordering of the items and results in treemaps with relatively high stability – that is, item locations do not change a great deal when the underlying data changes. A squarified treemap [24] (see Fig. 5b) arranges items from top-left to bottom-right sorted by value, and creates a layout with items with very low aspect ratios. Low aspect ratios are not only attractive from an aesthetic point of view, but treemaps where the items have low aspect ratios can also be expected to lead to lower Fitts' Law targeting times than treemaps that have items that are long and narrow. However, the squarified treemap can be unstable in very early stages of use or when application revisitation patterns change a lot over time. Because of their favourable performance in terms of stability and aspect ratio, the spiral and squarified treemaps are good candidates to be used in SCOTZ.

Even though previous work has demonstrated that the slight instability of the layout caused by size morphing (i.e., items unavoidably move as they grow/shrink) does not harm user performance [2], SCOTZ also offers a grid layout (see Fig. 5c) where no size morphing is applied, and is therefore very stable.

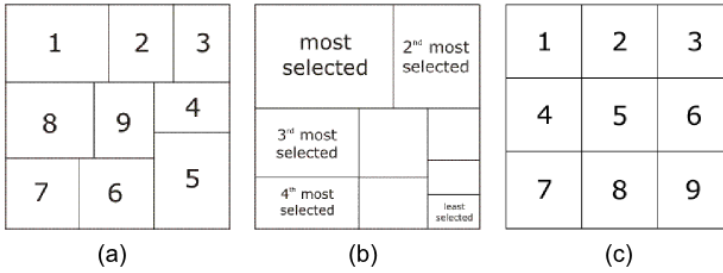


Fig. 5. Abstraction of the three layout options for application zones in SCOTZ: (a) spiral treemap, where applications are (arbitrarily) sorted once and this ordering is retained forever, and where zones are laid out in order in a clockwise spiral from top left; (b) squarified treemap, where application zones are sorted by size and laid out from top left to bottom right; (c) grid, where all application zones are of equal size and are laid out in a grid pattern.

The ordering of the window representations is either alphabetic by window title, by frequency (a combination of frequency and recency), or by the order in which the respective windows were opened, using a row-major order. Alphabetic ordering is very predictable and understandable; sorting by frequency supports window revisitation; and sorting the representations by the order in which the respective windows were opened is very stable when additional application windows are opened.

3.2 Improved Accessibility of Frequently Used Windows

SCOTZ allocates more space to the most frequently switched-to applications, which reduces the Fitts's Law [25] targeting time of these application zones, as well as making them easier to find [26].

3.3 Improved Accessibility of Recently Used Windows

When SCOTZ is bound to Alt+Tab, pressing Alt+Tab will bring up the SCOTZ interface as normal, but repeated presses of Tab will cycle through the windows according to z-order, similar to the implementation of Microsoft Windows Alt+Tab (see Fig. 6). This provides quick and easy access to recently used windows.

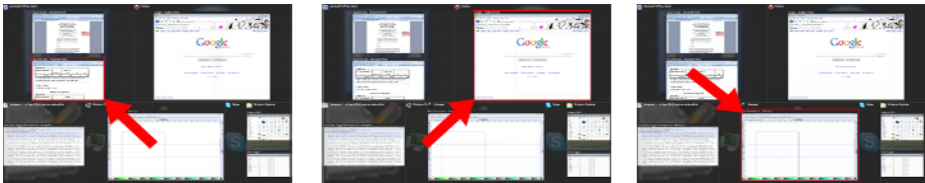


Fig. 6. SCOTZ Using Alt+Tab to cycle through the windows in SCOTZ by z-order. A pulsating red border indicates the currently selected window (indicated by a red arrow in this figure).

3.4 Support for Various Display Sizes

Modern display sizes range from very small (netbooks) to very large or multi-monitor setups. To accommodate this wide range of display sizes, SCOTZ can either be a full screen window (see Fig. 4), a smaller window in the centre of the screen, or a smaller window positioned under the mouse cursor (see Fig. 7). As the design of SCOTZ is aimed at keeping the application zones in spatially stable locations, positioning the SCOTZ interface relative to the mouse cursor means that a target can always be acquired with the same mouse gesture. Previous work has shown that despite some difficulties in the learning/training phase, mouse gestures can be very efficient and accurate [27].

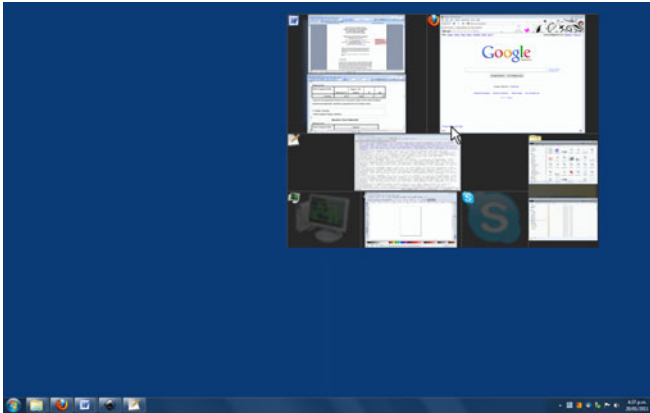


Fig. 7. SCOTZ in a smaller display mode, positioned under the mouse cursor

3.5 Support for Keyboard and Mouse Input

Previous work has shown that while most users prefer mouse-based methods for window switching, such as the Windows Taskbar, there is a small but significant subset of users who prefer keyboard-based methods, such as Alt+Tab [2]. Therefore, SCOTZ has rich options for both keyboard and mouse input.

3.6 Support for Application Launching

SCOTZ provides a single interface mechanism for both window switching and application launching. After having used SCOTZ for a long period of time, we expect users to be familiar with the locations of application zones, and if an application has no open windows associated with it, the application zone is still visible in SCOTZ (for example, see the Skype, Thunderbird and Window Media Player zones in Fig. 4). Therefore, it makes sense for these zones to double as efficient application launchers. Clicking on an empty application zone launches the application.

3.7 Options for Personalization

Personalization of an interface is driven by a variety of motivations [28], such as the personal goals of the user, accommodation for individual differences, or personal preference. Users can make several functional personalizations in SCOTZ to accommodate their personal goals (in addition to choices about layout and input methods, as presented in previous sections): for example, users can include or exclude certain applications in SCOTZ, and can customize the rate of growth for application zones. The appearance of SCOTZ is also customizable: users can change the font size and type to accommodate various levels of visual acuity, and the colour scheme to accommodate for various types of colour blindness as well as personal taste and preference. Also, users can adjust the opacity of SCOTZ (see Fig. 8).

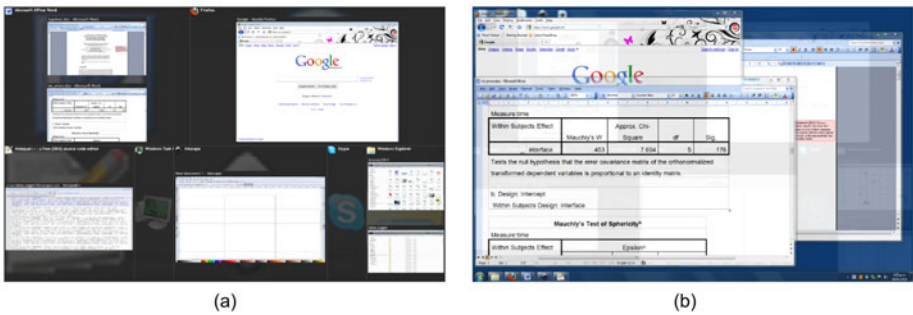


Fig. 8. Different opacity levels of the SCOTZ window, from (a) opaque to (b) almost transparent

4 Qualitative Study

A beta version of SCOTZ was given to five volunteers, who were asked to use it for at least several days. Next, they were given a questionnaire and were interviewed. Based on the results, we identified five main observations:

1. **People did not notice the slight location changes of application zones when using the spiral treemap layout.** Participants commented on never noticing *any* location changes of the application zones when the spiral treemap layout was used, even though gradual changes will have occurred as zones grew and shrunk.
2. **No clear preference for either the spiral or the squarified treemap layout.** The spiral layout was the default layout for SCOTZ, but some participants did (temporarily) switch to the squarified layout. However, there is no consensus on which layout is preferred. Some participants regarded the squarified layout as being too unstable (i.e., the application zones move too much), while others really liked the squarified layout and clearly preferred it over the spiral layout.
3. **Even users that mainly used Alt+Tab appreciated the size morphing of the application zones.** Though the size morphing of the application zones does not seem to have direct benefits for people that mainly use Alt+Tab for switching

between windows, participants commented that it helped them to guide their attention towards the application they were aiming for.

4. **The application launching functionality was only used by some participants, but it did not bother those who did not use it.** The option to use SCOTZ as an application launcher was only used by some users, yet this functionality did not bother non-users. If anything, retaining the application zones whilst no windows are open enhances the spatial stability of SCOTZ's layout.
5. **Overriding existing mappings such as Alt+Tab is useful, but risky.** Because SCOTZ was bound to Alt+Tab by default and SCOTZ retained Alt+Tab's (z-order) functionality, SCOTZ could be used without any additional learning. However, cycling to the correct application using SCOTZ (instead of clicking on the zones/thumbnails with the mouse) can be confusing, because it is harder to keep track of the selected item (also see Fig. 6). Possible solutions are (1) mapping SCOTZ under another key combination, (2) not retaining the z-ordering, but picking an ordering that matches the layout of SCOTZ better, or (3) providing better feedback on the z-ordering (e.g., with a small strip at the bottom of the screen showing the full order).

5 Lab Study

We performed a lab study to empirically compare the performance of SCOTZ, the Microsoft Windows 7 Taskbar, and Alt+Tab in a controlled environment. We chose the Taskbar and Alt+Tab for comparison because (1) Microsoft Windows is the most commonly used operating system, and therefore a comparison with the tools available in Windows 7 is relevant, and (2) these two tools present a *challenging* condition to compare SCOTZ against in terms of the key design principles of SCOTZ. The Windows 7 Taskbar places application icons in stable locations (unless a program is closed and re-opened), and Windows Alt+Tab provides explicit support for switching back to recently used windows because it places (the first six) window representations by their place in the z-ordering of windows, which is very similar to a recency ordering. We considered including Mac Os X Exposé as another comparison point, but excluded it because (although beautiful) its target acquisition time necessarily includes visual search time. In Exposé items do not appear in spatially stable locations; Exposé's representation of available windows is altered whenever windows are moved, opened, or closed, so users can never be certain where their target windows will appear. Several previous studies (including [2]) demonstrate that this type of spatial instability reduces target acquisition performance.

5.1 Method

In the experiment, participants were presented with a set of windows in Microsoft Windows 7, such as a word processor with several documents open, an e-mail application, a game, a video player, etc. Participants completed a series of tasks in which they had to switch to a particular window using the Windows 7 Taskbar, Windows 7 Alt+Tab, or SCOTZ in a successive series of tasks (see Fig. 9). In each condition, participants were instructed to use one particular switching tool *exclusively*.

Some windows were prompted often, while others were hardly ever prompted, following the findings reported in [2]: 80% of switches were to 35% of windows.

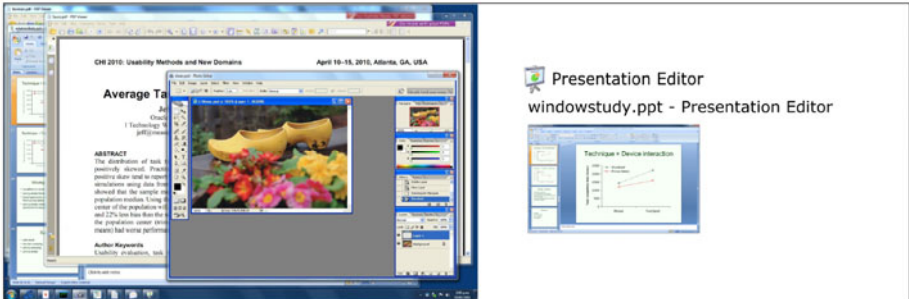


Fig. 9. The experimental interface: all windows are on the primary screen on the left, and the current task is on the secondary screen on the right

5.2 Design

Switching time and *errors* (switching to a non-target window) were measured across three levels of the independent variable *interface* (Taskbar, Alt+Tab and SCOTZ), and analysed using a one-way repeated-measures ANOVA. The experiment used a within-subject design and the order in which the conditions were presented to the participants was counterbalanced.

5.3 Procedure

At the start of each condition participants were given a verbal explanation and a demo of the window switching tool used in that particular condition. Participants then performed a series of 20 practice tasks with the window switcher before starting the experimental tasks.

At the start of each task, all windows were temporarily hidden from view and the user was prompted to press a 'Next' button at the centre of the secondary screen. Pressing this button revealed the next target window on the secondary screen (by showing the application icon, the window title and a window preview thumbnail of the target). Next, participants were prompted to click a 'Start' button in the centre of the primary screen, after which all the windows were unhidden, and participants then had to switch to the target window. If the participant switched to the incorrect window nothing happened. In total, participants performed 80 tasks in each condition (excluding the practice tasks).

After each condition, the participant filled out a short questionnaire regarding the window switching tool that had just been used.

5.4 Software and Hardware

All the content of the windows used in the experiment was non-modifiable to minimize distraction, to prevent accidental interaction with the windows, and to allow for

consistent window previews in the window switching tools throughout the tasks. To prevent learning effects across conditions, three different window sets were generated, all with unique applications and windows. The window sets were counterbalanced across the three conditions. Each window set contained 8 applications and 15 windows. For example, one of the window sets contained a PDF reader (with 4 windows open), a photo editor (3 windows), a presentation editor (2 windows), an HTML editor (2 windows), a music player, an email application, a command prompt, and a card game.

No window icons that are already in use by well-known applications were used, to ensure that all participants started off with equal knowledge about the application icons. This is particularly important for the evaluation of the Windows 7 Taskbar, which shows only application icons.

The full-screen version of SCOTZ using the squarified treemap algorithm was used. All application zones in SCOTZ were fixed (i.e., did not change during the experiment) and were pre-set to reflect the various frequencies with which applications/windows were switched to during the experiment.

A mouse with an extra side button was used in the experiment, and this side button was used to invoke SCOTZ.

5.5 Questionnaire

We used the NASA Task Load Index (NASA-TLX)² to assess perceived workload in each of the conditions. Two more questions were added to assess the perceived ease of learning to operate the window switcher (*operation*) and the perceived ease of learning window locations (*location learning*) in the window switcher. Also, participants were asked to rank the three window switching interfaces from most to least preferred.

5.6 Participants

Twelve people, all university students, participated in the experiment. Age ranged from 20 to 35 years old (mean 27); two participants were female. All participants were experienced computer users: computer use was at least 30 hours per week for each of the participants. Participants were reimbursed with a shopping voucher. The experiment took approximately 40 minutes to complete.

5.7 Results

Selection Times. We analysed the mean time to switch to a window for each of the methods (see Fig. 10). The results for the Windows 7 Taskbar are split by Taskbar button (for applications with only one associated window) and Taskbar thumbnail (for applications with more than one associated window, where the user first has to select the application icon and *then* the window in the fanned out sub-menu, see Fig. 1).

² <http://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX.pdf>

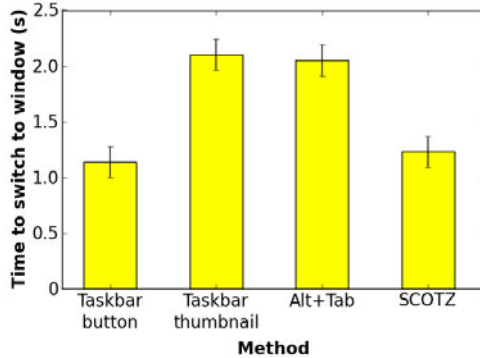


Fig. 10. Window selection times for the various methods. Error bars denote the 95% within-subject *CI* [29].

Mean window switching times when using a Taskbar button, a Taskbar thumbnail, Alt+Tab and SCOTZ are 1.1s, 2.1s, 2.1s and 1.2s, respectively, giving a significant effect of *interface*: ($F_{3,33}=53.3$, $p<.001$). Post hoc analysis (Bonferroni correction, $\alpha=.05$) reveals pairwise differences between all tools (all $p<.001$) except the Taskbar button and SCOTZ, and the Taskbar thumbnail and Alt+Tab.

By design, some of the target windows in the experiment were high up in the Alt+Tab ordering, and others further down, following a nearly uniform distribution across all possible positions. A detailed analysis of window switching times when Alt+Tab is used is shown in Fig. 11, which shows the selection times for Alt+Tab ranked by position of the target window in the Alt+Tab ordering, and split by input method (using the keyboard to sequentially step through the list of thumbnails, or using the mouse to click on the target thumbnail). Three observations are apparent from Fig. 11: (1) window selection time when using Alt+Tab with mouse input is relatively constant across positions of the target thumbnail, (2) window selection time when using Alt+Tab with keyboard input increases linearly as the position of the target thumbnail in the list of windows becomes higher ($r=.963$, $p<.01$), and (3) Alt+Tab with keyboard input is very efficient for switching back to the previously used window (position 1 in the ordering); the mean window switching time is 0.9 seconds for this particular type of window switch, which is shorter than the mean switching times for both the Taskbar and SCOTZ. However, this performance benefit quickly disappears when the target window is further down the list of windows.

Errors. Mean error rates for Taskbar button, Taskbar thumbnail, Alt+Tab and SCOTZ are 0.8%, 5.8%, 2.8% and 2.7%, respectively. The difference between these switching times is significant ($F_{3,33}=5.2$, $p<.01$). Post hoc analysis (Bonferroni correction, $\alpha=.05$) reveals a pairwise difference between the Taskbar button and Taskbar thumbnail ($p<.05$).

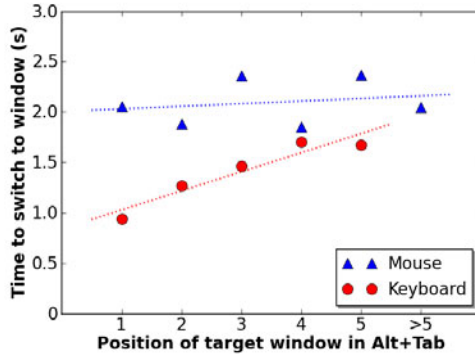


Fig. 11. Window selection times for Alt+Tab sorted by position of the target window in Alt+Tab and split by mouse and keyboard input. Participants almost never used the keyboard to select a window further than position 5 in the Alt+Tab ordering, hence there is no (reliable) data for this value.

Subjective Measures. The NASA-TLX worksheet results showed significantly different ratings for *mental demand*, *effort*, *location learning*, *frustration* (Friedman test, all $p < .001$), and *operation* ($p < .01$), also see Fig. 12. Post hoc pairwise comparisons (Bonferroni correction, $\alpha = .05$) reveal significant differences between Alt+Tab and SCOTZ on all five aforementioned factors, between the Taskbar and Alt+Tab on all these factors except *frustration*, and between the Taskbar and SCOTZ on the *mental demand* and *location learning* factors. All participants preferred SCOTZ the most, and 9 out of 12 participants preferred Alt+Tab the least (i.e., 3 out of 12 participants preferred the Taskbar the least).

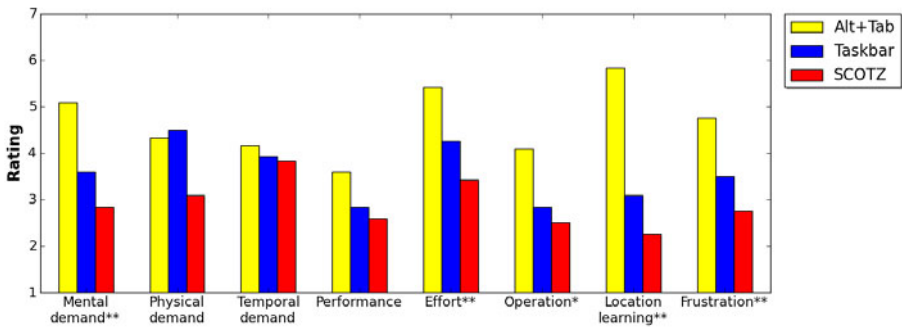


Fig. 12. Questionnaire results; lower ratings are better. * Difference is significant, $p < .01$. ** Difference is significant, $p < .001$.

5.8 Discussion

Window Switching Times. SCOTZ was faster than Taskbar thumbnails and Alt+Tab. Although there was no significant difference between SCOTZ and Taskbar buttons,

Taskbar buttons are not available when there is more than one window associated with the application (forcing users to resort to Taskbar thumbnails).

Subjective Measures. All participants ranked SCOTZ as the most preferred tool. Users perceived SCOTZ as less mentally demanding, costing less effort, and less frustrating than Alt+Tab, as well as finding it easier to learn to operate and to learn item locations in SCOTZ compared to Alt+Tab. Also, window locations in SCOTZ were perceived as easier to learn than those on Taskbar, and SCOTZ was perceived as less mentally demanding than the Taskbar. These two factors are likely related: SCOTZ is less mentally demanding because it is easier to learn locations, thereby reducing the cognitive burden for users. It is interesting that users found locations in SCOTZ easier to learn than locations of items on the Taskbar, as in both cases these were completely stable in the current study.

Alt+Tab. Overall, Alt+Tab was unpopular, with 75% of participants ranking it as least preferred. Alt+Tab was also judged to be more mentally demanding and costing more effort than the Windows Taskbar. Last, users found it harder to learn to operate and learn item locations in Alt+Tab than with the Taskbar. One participant commented that he/she “hated how Alt+Tab icons moved around”. However, these results for Alt+Tab might have been negatively influenced by the fact that users had to use Alt+Tab for *all* window switches in the experiment. Our results show that Alt+Tab is very efficient for switching back to the most recently used window, with this particular type of switch outperforming both the Taskbar and SCOTZ. One participant commented “I use Alt+Tab to switch between the most recent windows, and other methods for older windows.” Such a ‘mixed approach’, i.e. using Alt+Tab to switch back to the most recently used window, but another method for other types of window switches might lead to higher user satisfaction than the ‘enforced’ use of Alt+Tab for all window switches that was the case in the experiment.

Comparison to other window switching tools. Our experiment compared user performance with SCOTZ against that with the Windows 7 Taskbar and Alt+Tab. Further work is needed to compare SCOTZ performance with that of the wide range of research and commercial tools reviewed in Section 2. However, we believe SCOTZ’s key design goals – supporting spatially stable means for acquiring windows and applications, and providing support for rapidly retrieving frequently and recently retrieved windows – are important for enabling high performance window acquisition. In particular, inconstant spatial locations are likely to force users to resort to time consuming visual search (to seek a target) or decision making (to calculate the effect of an algorithm, for example).

6 Conclusions and Future Work

While previous work has found that window revisitation is very common, no tools developed so far explicitly support this revisitation. We used this finding to inform the design of a new window switcher called SCOTZ, which supports window revisita-

tion by increasing the size of the most switched-to applications, and keeping them in positions that are as stable as possible.

Our lab study demonstrates the performance benefits of SCOTZ over two common window switching tools: the Microsoft Windows 7 Taskbar and Alt+Tab. This study also generated valuable insights regarding the most recent window switching tools available in Microsoft Windows 7.

More research into the suitability of SCOTZ for Alt+Tab users could shed more light on how these users can best be supported in their window switching activities. Interestingly, even Alt+Tab users reported benefits from the size morphing of the application zones in SCOTZ, but retaining the Alt+Tab order in SCOTZ did confuse these users. Ideally, SCOTZ should retain the rapid back-and-forth switching between two windows that Alt+Tab offers (see results of the lab study) while also assisting users in finding items further down the Alt+Tab ordering.

Finally, we are developing a new treemap algorithm to better suit SCOTZ than existing algorithms, in terms of enhanced spatial stability of the application zones.

References

1. Hutchings, D.R., Stasko, J.: Revisiting display space management: understanding current practice to inform next-generation design. In: Proc. of GI 2004, pp. 127–134. Canadian Human-Computer Communications Society (2004)
2. Tak, S., Cockburn, A., Humm, K., Ahlström, D., Gutwin, C., Scarr, J.: Improving window switching interfaces. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5727, pp. 187–200. Springer, Heidelberg (2009)
3. Bannon, L., Cypher, A., Greenspan, S., Monty, M.L.: Evaluation and analysis of users' activity organization. In: Proc. of CHI 1983, pp. 54–57. ACM Press, New York (1983)
4. Henderson, D.A., Card, S.: Rooms: the use of multiple virtual workspaces to reduce space contention in a window-based graphical user interface. *ACM Trans. Graph.* 5(3), 211–243 (1986)
5. Smith, G., Baudisch, P., Robertson, G., Czerwinski, M., Meyers, B., Robbins, D., Horvitz, E., Andrews, D.: Groupbar: The taskbar evolved. In: Proc. of OzCHI 2003, pp. 34–43 (2003)
6. Bardram, J., Bunde-Pedersen, J., Soegaard, M.: Support for activity-based computing in a personal computing operating system. In: Proc. of CHI 2006, pp. 211–220. ACM Press, New York (2006)
7. Robertson, G., Horvitz, E., Czerwinski, M., Baudisch, P., Hutchings, D., Meyers, B., Robbins, D., Smith, G.: Scalable fabric: Flexible task management. In: Proc. of AVI 2004, pp. 85–89. ACM Press, New York (2004)
8. Robertson, G., van Dantzig, M., Czerwinski, M., Hinckley, K., Thiel, D., Robbins, D., Ridsen, K., Gorokhovskiy, V.: The task gallery: A 3D window manager. In: Proc. of CHI 2000, pp. 494–501 (2000)
9. Kaptelinin, V.: Umea: translating interaction histories into project contexts. In: Proc. of CHI 2003, pp. 353–360. ACM, New York (2003)
10. Dragunov, A.N., Dietterich, T.G., Johnsrude, K., McLaughlin, M., Li, L., Herlocker, J.L.: Tasktracer: a desktop environment to support multi-tasking knowledge workers. In: Proc. of IUI 2005, pp. 75–82. ACM, New York (2005)

11. Oliver, N., Smith, G., Thakkar, C., Surendran, A.: Swish: Semantic analysis of window titles and switching history. In: Proc. of IUI 2006, pp. 194–201. ACM Press, New York (2006)
12. Oliver, N., Czerwinski, M., Smith, G., Roomp, K.: Relalttab: assisting users in switching windows. In: Proc. of IUI 2008, pp. 385–388. ACM, New York (2008)
13. Xu, Q., Casiez, G.: Push-and-pull switching: window switching based on window overlapping. In: Proc. of CHI 2010, pp. 1335–1338. ACM, New York (2010)
14. Shneiderman, B.: Direct manipulation for comprehensible, predictable and controllable user interfaces. In: Proc. of IUI 1997, pp. 33–39. ACM, New York (1997)
15. Tashman, C.: Windowscape: A task oriented window manager. In: Proc. of UIST 2006, pp. 77–80. ACM Press, New York (2006)
16. Bernstein, M., Shrager, J., Winograd, T.: Taskposé: exploring fluid boundaries in an associative window visualization. In: Proc. of UIST 2008, pp. 231–234. ACM, New York (2008)
17. Gaylin, K.B.: How are windows used? Some notes on creating an empirically-based windowing benchmark task. In: Proc. of CHI 1986, pp. 96–100 (1986)
18. Hutchings, D., Smith, G., Meyers, B., Czerwinski, M., Robertson, G.: Display space usage and window management operation comparisons between single monitor and multiple monitor users. In: Proc. of AVI 2004, pp. 32–39. ACM Press, New York (2004)
19. de Chiara, R., Erra, U., Scarano, V.: A visual adaptive interface to file systems. In: Proc. of AVI 2004, pp. 366–369. ACM, New York (2004)
20. Grudin, J.: Partitioning digital worlds: Focal and peripheral awareness in multiple monitor use. In: Proc. of CHI 2001, pp. 458–465 (2001)
21. Kumar, M., Paepcke, A., Winograd, T.: Eyeexpose: Switching applications with your eyes. Tech. rep. (2007)
22. Shneiderman, B.: Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.* 11(1), 92–99 (1992)
23. Tu, Y., Shen, H.: Visualizing changes of hierarchical data using treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 1286–1293 (2007)
24. Bruls, M., Huizing, K., Wijk, J.v.: Squarified treemaps. In: Proceedings of Joint Eurographics and IEEE, pp. 33–42. IEEE Press, Los Alamitos (2000)
25. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47, 381–391 (1954)
26. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience* 5(6), 495–501 (2004)
27. Dulberg, M.S., Amant, R., Zettlemoyer, L.: An imprecise mouse gesture for the fast activation of controls. In: Proc. of INTERACT 1999, pp. 375–382 (1999)
28. Blom, J.: Personalization: a taxonomy. In: Proc. of CHI 2000 Extended Abstracts, pp. 313–314. ACM, New York (2000)
29. Masson, M.E.J., Loftus, G.R.: Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology* 57(3), 203–220 (2003)

Evaluating Commonsense Knowledge with a Computer Game

Juan F. Mancilla-Caceres and Eyal Amir

Computer Science Department,
University of Illinois at Urbana-Champaign,
201 N. Goodwin Avenue, Urbana, IL 61801, USA
{mancill11,eyal}@illinois.edu

Abstract. Collecting commonsense knowledge from freely available text can reduce the cost and effort of creating large knowledge bases. For the acquired knowledge to be useful, we must ensure that it is correct, and that it carries information about its relevance and about the context in which it can be considered commonsense. In this paper, we design, and evaluate an online game that classifies, using the input from players, text extracted from the web as either commonsense knowledge, domain-specific knowledge, or nonsense. A continuous scale is defined to classify the knowledge as nonsense or commonsense and it is later used during the evaluation of the data to identify which knowledge is reliable and which one needs further qualification. When comparing our results to other similar knowledge acquisition systems, our game performs better with respect to coverage, redundancy, and reliability of the commonsense acquired.

1 Introduction

A vast amount of information about the world is needed to create an artificial commonsensical agent [7]. This kind of knowledge, which includes facts about events, and objects, is what we call commonsense knowledge. Collecting commonsense knowledge is difficult because it is dynamic and dependent on context. This makes it impossible to generate it randomly and to verify it automatically, which implies that humans are needed to either collect the commonsense knowledge or to verify it.

The main difficulty of needing humans is that they need to be encouraged to participate. As previous studies show [2], contributors tend to produce noisy data even when they are paid for their services because they are trying to maximize the reward, whether they are producing the right data or not.

In this paper we introduce a game that takes text extracted automatically from the Web and uses input from players to verify it as commonsense. The main difficulty of this approach lies in the fact that the game needs to encourage players to supply the correct information. With the help of the players, our game classifies the knowledge as commonsense, domain-specific, or meaningless. It also reports if more information about a given fact is needed in order to classify it correctly. Correctness of the data is ensured through the design of several stages within the game, and through restricting communication among players. Also, we create a continuous scale that ranges from nonsensical (or unknown) sentences to commonly known facts. A discretization of

such scale can be used to classify the sentences as commonsense or not. This scale also allows us to clearly identify which knowledge needs revision.

The focus of this paper is to present a method that provides guarantees on the correctness of the data collected through a game. Although the design of the game does not constrain the source of the input data, we are currently obtaining it from *Simple Wikipedia* [13] because some of its policies regarding the content of the articles are appropriate for commonsense extraction. For example, unverified research is not allowed in any article. Within a period of five weeks, more than 150 people played the game and more than 3,000 sentences were evaluated.

2 Related Work

The best known approaches to gather commonsense are Cyc [5] and OpenMind [11]. Cyc uses experts to input the knowledge, whereas OpenMind used volunteers. Some issues with these approaches are that Cyc requires the user to be familiar with their language, and OpenMind lacks a way to motivate volunteers to participate.

Another effort to collect common knowledge from contributors is LEARNER2 [1]. It collected data about *part-of* relations. Unlike our game, LEARNER2 lacks the capability of redirecting the user's effort to improve the reliability of data already collected. There are also several games that encourage participation of users to enter data into knowledge bases. Among these, Cyc released the game FACTory which is similar in format to the first stage of our game but does not include any way to guarantee the correct behavior of players. *Verbosity* [12], uses two players to fill in templates, and *Common Consensus* [6] asks two players questions about achieving a given goal.

Combining the computational power of machines and humans has been addressed in [10]. The use of a continuous scale to reason about commonsense knowledge was explored before in [3].

3 Game Design

The initial input information for the game comes from an off-the-shelf parser [4] that extracts a sentence from an article in *Simple Wikipedia* and, together with the action of the user, produces an update to the knowledge base as the output.

The simplest implementation of the game would have a single user classifying sentences as either commonsense or not. This is not enough because it would be impossible to evaluate the answers of the player as correct or incorrect. Also having several players evaluating the same sentence and accepting the input only if they agree amongst one another is not appropriate because it would be easy for a group of players to act in collusion and agree on entering the same answer, regardless of the question.

The basic problem is that human players can always agree on a fixed strategy, and yes/no questions are not enough to correctly classify the knowledge. To solve this, we add a non-human player to the set of players and classify the input text in four different categories: Nonsense, Unknown, True, False.¹

¹ If the parser extracted an incomplete sentence or any other nonsensical data, the sentence is nonsense, if the sentence was extracted correctly the content may either be known (true or false), or unknown.

We devised a three player game in which the purpose of the player is to distinguish between another human and a machine. Both humans will give the same answer on the sentence while the machine will guess its answer. The design of the game solves the problem outlined before: The two humans can no longer agree on any strategy because the identity of the players is unknown. Also, if the answer of one player does not follow commonsense, the other human might erroneously identify the player as a machine, which results in a penalization on the player's score.

Because commonsense depends on the context, it is necessary to consider context explicitly. In [8], the author proposes a formula $Holds(p,c)$ to assert that the proposition p holds in context c . Using this idea, the appropriate task for the player is to answer a question based on that formula. In our case, the context is handled by the name of the *Simple Wikipedia* article used as source for the sentence. The context can be used by the player to answer correctly, while addressing the problem of uninstantiated sentences that may be produced by the parser.

Figure 1 shows snapshots of the game in all its stages. The game works as follows:

- In the first stage, the player chooses a topic (which matches the title of the Wikipedia article from which a sentence is to be retrieved).
- A sentence is randomly selected from the article. The system chooses either a new sentence or a sentence that has been verified before. This balances the coverage and reliability of the data by increasing the times a sentence has been verified. Then, the player indicates whether the fact expressed by the sentence is true in the context of the article using the four options previously described.

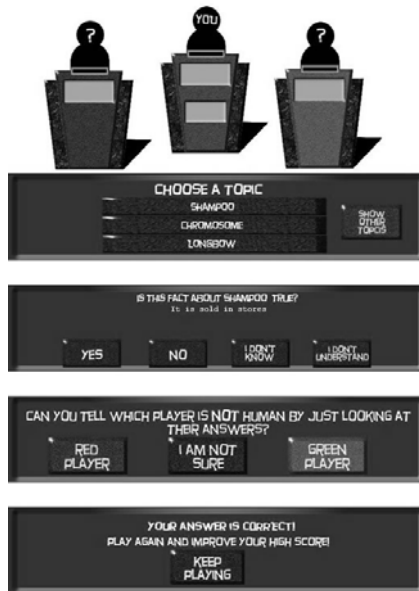


Fig. 1. Snapshot of the game. All the stages of the game are shown one below the other.

- In the second stage, the player sees the answer of the other two players and identifies which of the two is the machine that is answering randomly. In the case of a single player playing the game, the other answer comes from recorded games. If it is impossible to distinguish between the two players, there is an option to pass and avoid making a decision. If the player identifies the human as the machine, points are deducted; otherwise, points are awarded.
- After this, the player gets the opportunity to play again.

4 Identifying Commonsense Knowledge

A majority vote is not enough to identify a fact as commonsense, because we have more confidence if a sentence was evaluated by a large amount of players rather than by a few. Thus, we create a scale of commonsense that describes how common a specific fact is. The scale needs to be proportional to the ratio of people who know the given fact, and also contain information about the confidence of such ratio. We first define four quantities, t_{count} , f_{count} , u_{count} , n_{count} , that hold the number of times a sentence s has been classified as true, false, unknown, and nonsense, respectively.

Definition 1. Let $P_\sigma(s)$ be the ratio of people that have answered true, false, and unknown over the total number of instances the sentence s has been verified. Let $m(s) = t_{count}(s) + f_{count}(s) + u_{count}(s) + n_{count}(s)$.

$$P_\sigma(s) = \frac{t_{count}(s) + f_{count}(s) + u_{count}(s)}{m(s)}. \quad (1)$$

Under the assumption of independence, each instance of the game can be considered a Bernoulli trial. $P_\sigma(s)$ is then an estimator of the real proportion of people that understand the sentence. Our null hypothesis is that the ratio of people classifying the sentence as nonsense should be 0.5. If we fail to reject the null hypothesis, we must conclude that we don't have enough information to identify the sentence as meaningful or nonsense.

Definition 2. Let $e_n(s)$ be the *effect size*, the difference between the actual and expected number of times the sentences have been marked as nonsense.

$$e_n(s) = \left| n_{count}(s) - \frac{m(s)}{2} \right|. \quad (2)$$

Definition 3. Let $p_n(s)$ be the p-value of the Binomial Hypothesis Test, the probability of observing a difference in the value of a random variable of at least the size of the effect size $e_n(s)$.

$$p_n(s) = P\left(X < \frac{m(s)}{2} - e_n(s)\right) + P\left(X > \frac{m(s)}{2} + e_n(s)\right). \quad (3)$$

The p-value $p_n(s)$ is the probability of observing the current counters given the null hypothesis. The lower its value, the more confident we are about their values.

Table 1 shows some sentences with their corresponding $P_\sigma(s)$ and $p_n(s)$. Notice that $P_\sigma(s)$ and the p-value cannot distinguish amongst all sentences because one only considers the ratio of people that agree on the sentence, whereas the other only considers the amount of people that has evaluated the sentence. With this in mind we define $\pi_s(s)$, which allows us to easily classify sentences as meaningful or nonsense.

Table 1. The Id is used to refer to each sentence in the paper. Eval refers to the number of times that the sentence has been evaluated. $P_\sigma(s)$ and $P_\gamma(s)$ are the proportion of people who didn't answer *nonsense*, or who answer *true* or *false*, respectively. The value of $\pi_s(s)$ and $\pi_c(s)$ represents the confidence that we have when classifying the sentence as meaningful or known, respectively. The last column is the decision made with a significance of 0.1.

Id	Sentence	Article	Eval	$P_\sigma(s)$	p-value	$\pi_s(s)$	Meaningful
1	<i>People are known acting in comedies are comedians</i>	Comedy	1	1	1	0.5	Unknown
2	<i>Computer can use many bits</i>	Computer	6	1	0.03	0.98	Yes
3	<i>For example some languages (e.g.Chinese,Indonesian)</i>	Verb	6	0.17	0.03	0.02	No
Id	Sentence	Article	Eval	$P_\gamma(s)$	p-value	$\pi_c(s)$	Known
4	<i>It is a county in the U.S. state of North Carolina</i>	Anson County	9	0	0.004	0.002	No
5	<i>The level experience is needed to level</i>	Diablo II	1	1	1	0.5	Unable
6	<i>Chess is a very complex</i>	Chess	9	1	0.004	0.99	Yes

Definition 4. Let $\pi_s(s)$ be the value that represents how much confidence we have on a sentence s being meaningful.

$$\pi_s(s) = \begin{cases} 1 - p_n(s)/2 & \text{if } P_\sigma(s) > 0.5 \\ p_n(s)/2 & \text{if } P_\sigma(s) \leq 0.5 \end{cases} \tag{4}$$

To classify the sentence as meaningful, we only need to define a threshold α against which we can compare $\pi_s(s)$. If $\pi_s(s) < \alpha$ we have a confidence of $1-\alpha$ that the sentence is nonsense and if $\pi_s(s) > 1-\alpha$, we have a confidence of $1-\alpha$ that the sentence is meaningful. Otherwise, we can only conclude that we need more players to evaluate the sentence. We perform a similar analysis to the one described previously to define a scale $\pi_c(s)$ that represents the fact that a given sentence s is commonly known. In order to classify a sentence as commonsense we combine both $\pi_s(s)$ and $\pi_c(s)$.

Definition 5. Let $\pi(s)$ represent the confidence about a sentence being commonsense.

$$\pi(s) = \pi_s(s)\pi_c(s) \tag{5}$$

Table 2 shows the corresponding value of $\pi(s)$ of the sentences from Table 1. Notice that to classify a sentence as commonsense it requires both $\pi_s(s)$ and $\pi_c(s)$ to be high.

Table 2. Id, Eval, $\pi_s(s)$, and $\pi_c(s)$ are defined as in Table 1. $\pi(s)$ represents the confidence that we have on identifying each sentence as commonsense.

Id	Eval.	$\pi_s(s)$	$\pi_c(s)$	$\Pi(s)$	Commonsense
1	1	0.5	0.5	0.25	Unknown
2	6	0.98	0.98	0.97	Yes
3	6	0.02	0.5	0.01	No
4	9	0.99	0.002	0.002	Domain-specific
5	8	0.004	0.5	0.002	No
6	9	0.99	0.99	0.99	Yes

5 Evaluation

Coverage, reliability, and identifying the presence of knowledge that needs further classification are of primary interest to knowledge acquisition systems, especially when the knowledge comes from volunteer contributors. In contrast to other systems, our game offers an explicit way to detect knowledge that should be discarded due to errors or noise in the input of contributors. Also, all other previous games do not provide any way to distinguish the data that need further qualification. These features are achieved by the use of our scale $\pi(s)$. If a sentence is not nonsense, commonsense or domain-specific, then the game can be directed to present it to players more often until enough data has been collected to make a decision regarding such sentence.

Among the reviewed systems, only LEARNER2 reports data about redundancy. Out of 6658 entries, only 2088 are different statements and 4416 entries yielded only 350 distinct statements. This means that they collected 1.29 entries per statement. These few entries per statement produce unreliable data, which means that only 350 statements can actually be trusted. In contrast, our game collected 6763 entries and generated 3011 evaluated sentences, with an average of 3.46 entries per statement. Therefore, our data is more reliable than that of LEARNER2. Figure 2 shows the comparison of coverage and reliability between LEARNER2 and our game.

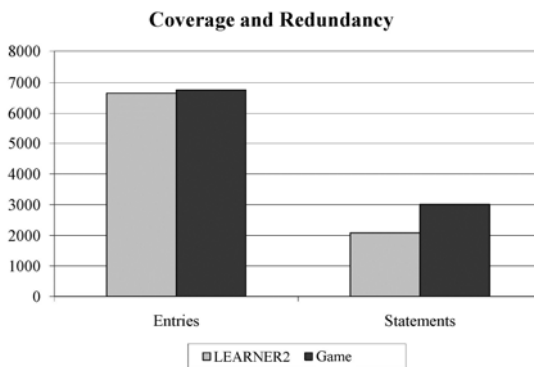


Fig. 2. Comparison between LEARNER2 and our game

For the evaluation, we asked 4 judges to classify a random sample of 50 sentences from our knowledge base. The judges evaluated the knowledge by classifying it in these categories: "*Generally/Definitively True*", "*Sometimes/Probably True*", "*Unknown*" and "*Nonsense/Incomplete*", which correspond to Commonsense, Domain-Specific, Unknown, and Nonsense, respectively. This categories are similar to the ones used by [9]. When comparing the answers of the judges to the ones from the game, the average agreement between players and judges was 94% ($\alpha=0.1$).

In comparison to the other systems, *Verbosity* asked the judges to rate each input as correct or incorrect; the judges reported 0.85 of the data to be correct. LEARNER2 used a scale similar to ours and reported that 89.8% of the data that was entered by at least 2 people was correctly common knowledge. Our game outperforms the previous systems.

6 Conclusions and Future Work

We presented the design of a game that evaluates and classifies sentences extracted automatically from the Web. The main advantage of our design is that it classifies commonsense knowledge in a continuous scale, which allows us to talk about how common a commonsense fact is. Our analysis gives us confidence about the results even when some of the players disregard the rules and create noisy data. Also, we distinguish between data that needs to be evaluated further and data that has been classified with certainty. Although the game has already provided data that shows that our approach is viable, improvements in the design of the game are possible. One feature that will be further explored in future work is the demographics of the players. Each answer given by the player is stored according to their age group and location. This is useful because we will not only be able to classify commonsense knowledge, but we will also be able to cluster commonsense knowledge according to demographic information.

References

1. Chlovski, T., Gil, Y.: An analysis of knowledge collected from volunteer contributors. In: AAAI 2005: Proceedings of the 20th National Conference on Artificial Intelligence, pp. 564–570. AAAI Press, Menlo Park (2005)
2. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are your participants gaming the system?: screening mechanical turk workers. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, pp. 2399–2402. ACM, New York (2010)
3. Druzdzal, M.J.: Probabilistic reasoning in decision support systems: from computation to common sense. PhD thesis, Pittsburgh, PA, USA (1993)
4. Hadidi, B., Johri, N., Pantley, D., Pradham, A., Wang, F.: Automated knowledge extraction from wikipedia. Available upon request to authors (2010)
5. Lenat, D.B., Guha, R.V., Pittman, K., Pratt, D., Sheperd, M.: Cyc: toward programs with common sense. *Commun. ACM* 33(8), 30–49 (1990)

6. Lieberman, H., Smith, D.A., Teeters, A.: Common Consensus: a web-based game for collecting commonsense goals. In: Proceedings of the Workshop on Common Sense and Intelligent User Interfaces held in conjunction with the 2007 International Conference on Intelligent User Interfaces (IUI 2007) (2007)
7. McCarthy, J.: Programs with common sense. In: Semantic Information Processing, pp. 403–418. MIT Press, Cambridge (1968)
8. McCarthy, J.: Artificial intelligence, logic and formalizing common sense. In: Philosophical Logic and Artificial Intelligence, pp. 161–190. Kluwer Academic, Dordrecht (1990)
9. Schubert, L., Tong, M.: Extracting and evaluating general world knowledge from the brown corpus. In: Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning, pp. 7–13. Association for Computational Linguistics, Morristown (2003)
10. Shahaf, D., Amir, E.: Towards a theory of AI completeness. In: 8th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 2007) (2007)
11. Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T., Zhu, W.L.: Open mind common sense: Knowledge acquisition from the general public, pp. 1223–1237. Springer, Heidelberg (2002)
12. von Ahn, L., Kedia, M., Blum, M.: Verbosity: a game for collecting commonsense facts. In: Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems. Games, vol. 1, pp. 75–78. ACM Press, New York (2006)
13. Simple Wikipedia, <http://simple.wikipedia.org>

Remote Usability Testing Using Eyetracking

Piotr Chynał and Jerzy M. Szymański

Institute of Informatics, Wrocław University of Technology
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

Piotr.Chynał@pwr.wroc.pl, 157690@student.pwr.wroc.pl

Abstract. In the paper we present a low cost method of using eyetracking to perform remote usability tests on users. Remote usability testing enables to test users in their natural environment. Eyetracking is one of the most popular techniques for usability testing in the laboratory environment. We decided to try to use this technique in remote tests. We used standard web camera with freeware software. Our experiment showed that such method is not perfect, but it could be a good addition to the standard remote tests, and a foundation for further development.

Keywords: Eyetracking, Usability, Remote Usability Testing, Human-Computer Interaction.

1 Introduction

In modern usability testing, remote tests are becoming more and more popular. Because of lack of time and money, companies are looking for alternatives for standard tests with users. Those tests require a place to test, gathered users, moderator and equipment. This situation lead to idea of remote usability testing. Its main goal is to test users in their natural working place, without any sophisticated equipment. Users in their own environment are behaving more naturally, like they would normally do while using the given website [4]. Moreover we do not need to gather all the users at one time, we can work with them when they have the time to take part in the test. Also we can have participants from different cities or even countries that would normally not visit our laboratory. Furthermore comparisons of the results of standard laboratory and remote testing have shown that participants find the same usability issues on tested pages with both methods [2], [6], [9].

So far remote usability tests are evaluated using standard methods such as remote surveys and video conferences [1], [7]. We can also use some traditional laboratory usability testing methods in remote environment [8]. One of the most popular tool for standard usability testing is eyetracking [3], [5]. The biggest drawback of eyetracking technology is its cost. The equipment is very expensive and companies which perform such tests usually charge a lot of money for such tests. We tried to perform remote usability test with eyetracking, using low cost hardware, such as ordinary web camera and free software.

2 Tools Used in Experiment

The main equipment used during this research was a casual web cam Logitech Quick Cam Pro 9000. Software used in this experiment was:

1. Piotr Zieliński Opengazer¹ .Net port made by Przemysław Nibyłowicz². This application enables gazetracking using ordinary webcam. It is freeware open source software. After selecting feature face points on the video image, user calibrates the program by looking at the appearing squares. Next, when calibration is finished, line of gaze is tracked by the program. We slightly modified this application, so it stores the data of all the points that users is looking at in the text file (Fig. 1).

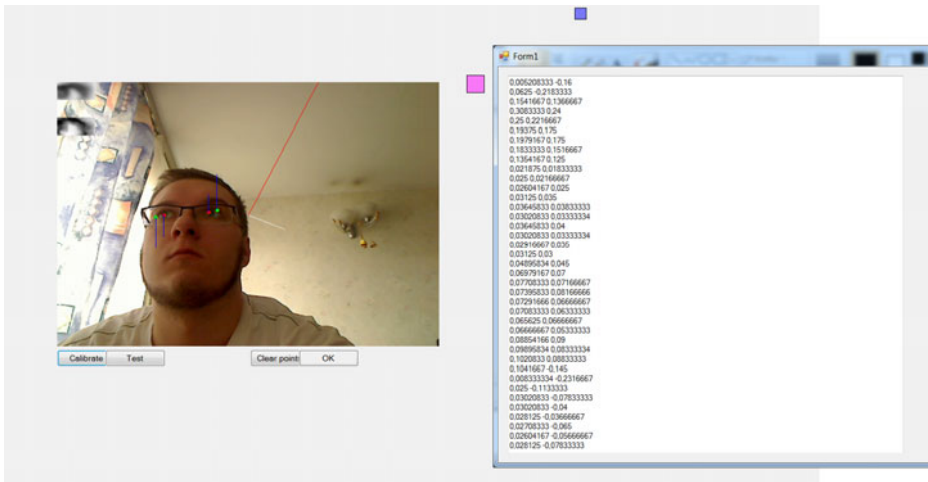


Fig. 1. Application after the calibration process. Blue square shows where the participant is looking and in the background we have a window that shows the coordinates of the gaze.

The points in the application have coordinates from -1 to 1 (float variable), so to store them as actual points on the screen, we are transforming them as shown below:

```
int okoX = (int)((EventArgs.EyeX + 1.0f) * (w / 2));
```

```
int okoY = (int)((EventArgs.EyeY + 1.0f) * (h / 2));
```

We add 1.0 to the float value (EventArgs.EyeX or EventArgs.EyeY) and then we multiply this value by height (h) or width (w) of the screen divided by two. Then the value is parsed to integer. Furthermore we needed to transform those points so they could be used in the heatmap generator application (point 2 below). We needed to transform the screen parameters to image parameters:

¹ <http://www.inference.phy.cam.ac.uk/opengazer/>

² <http://netgazer.sourceforge.net>

```
int imageX = (int) ((1024*okoX)/w);
int imageY = (int) ((768*okoY)/h);
```

We received new coordinates for the image by calculating the given point with the image size and screen size.

- JavaScript application for creating heatmaps³. Created by Michael Dungan, released under the MIT license. This application creates heatmaps on images taking mouse movement as an input data (Fig. 2). However it has the functionality to import points, so we used it to generate heatmaps for our test. It takes as parameters the offset of the picture and a “mousemove mask” for smooth rendering, so we omitted this last parameter and put ‘3’ as a default value (center). Additionally we needed to run the browser in full screen mode, so the coordinates of the points calculated in Opengazer .Net port application would be adequate.

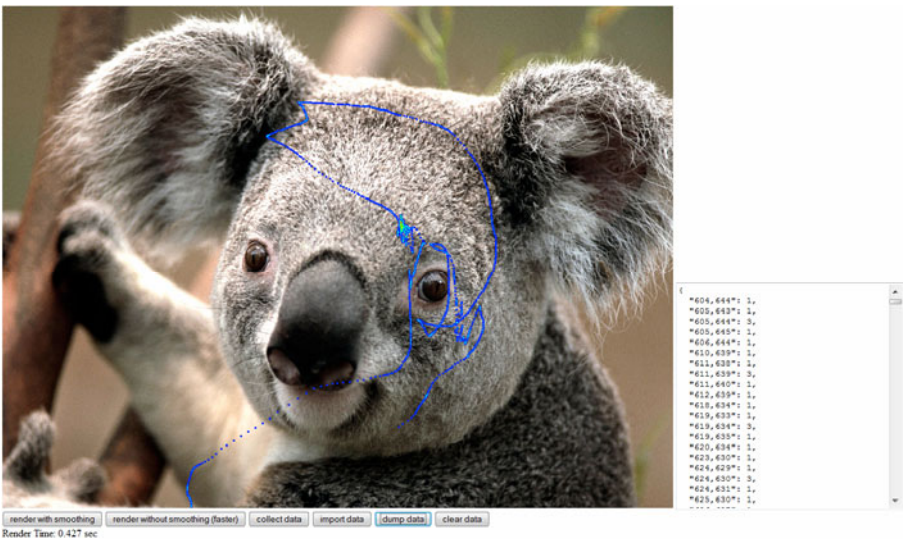


Fig. 2. Application for creating heatmaps from mouse input as well as from imported points (Source for koala bear picture: Microsoft Windows 7 sample images)

- Real VNC⁴ – simple application for desktop sharing.
- Skype⁵ – popular application that enables talking through microphone over the internet.
- CamStudio⁶ – simple application for recording actions from computer screen.

³ <https://github.com/xxx/heatmap>

⁴ <http://www.realvnc.com>

⁵ <http://www.skype.com>

⁶ <http://www.camstudio.org>

3 Experiment

The goal of our experiment was to try to perform remote usability eyetracking test and analyze if our method is suitable for future development. We tested it with five users, on different computers but with the same webcam (Logitech Quick Cam 9000). During the test, first thing that we did was to connect to user's computer via Skype and Remote VNC. After that we instructed the user how to position the camera and how to calibrate the eyetracking application. This was the toughest part for all the users, because the calibration process in this application requires a lot of patience. Application sometimes crashes, and calibration often needs to be repeated few times until it is correctly set. When the results of calibration were satisfactory we asked our user to start CamStudio and perform some simple actions on google.com web page, such as to log in. While user was working with the page we were able to see where he is looking, because small blue square was showing that position (Fig. 3). It is a very helpful thing for the moderator, because he can observe at which elements the user is looking during the test and have more control over what user is doing.



Fig. 3. Google.com website during the test, small blue square shows the position of the participants gaze

After the experiment we took the coordinates of the user's gaze on the screen (stored in a text file) and we imported them to the heatmap application. It allowed us to create some simple heatmaps for our test (Fig. 4).

To sum up the experiment, our method of remote usability eyetracking test can provide the moderator with:

1. Verbal remarks and comments of the tested users, as in standard remote test.
2. Observation of user actions and where user was looking during the test via Remote VNC.
3. Recordings from the point 2 provided by CamStudio.
4. Text file with the coordinates of the points on the screen on which the user was looking during the test. Those points can be used to generate some visual reports.

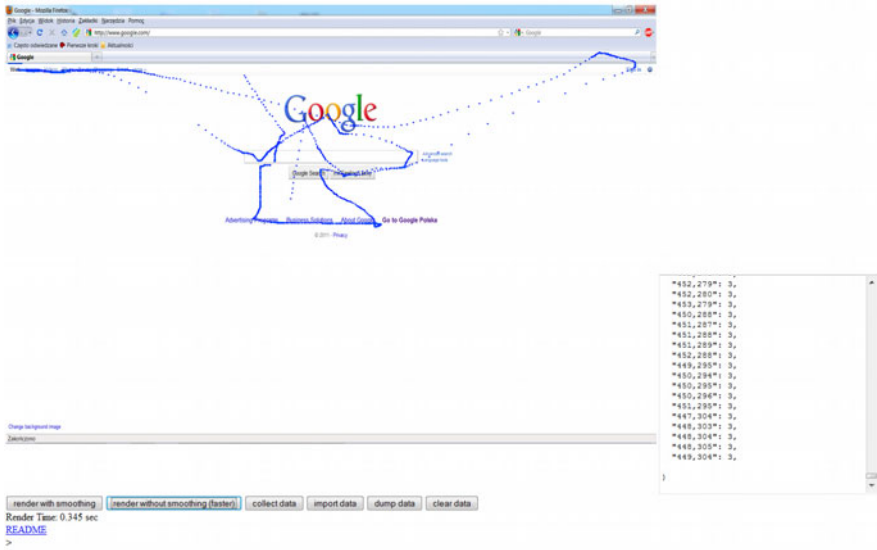


Fig. 4. Simple heatmap for google.com website created with JavaScript heatmap generator

4 Conclusions and Future Work

Our experiment has shown that it is possible to perform remote usability eyetracking tests. Moreover we obtained a lot of valuable data that could be processed for usability analysis. However our solution has some drawbacks that need to be addresses. First of all user has to perform many actions before the test, such as setting up the camera and calibrating the eyetracking software. In future we will try to improve this process, so the users will not need to perform so many operations. Secondly, using a web cam and Opengazer .Net port we need to perform calibration many times to obtain eyetracking data, which precision can be compared with professional eyetrackers in the laboratory environment. So far we managed to obtain the precision which is about two times worse than in professional eyetrackers, so there is still room for further development here. We will also try to improve calibration process, so it would be faster and more effective. Last thing is that we needed to put a lot of effort into creating a single heatmap for our test. We need to create an application that could quickly transform received points into heatmaps, gaze plots and other useful reports.

In conclusion our method needs a lot of improvements, but it is definitely a good starting point for creating a remote usability eyetracking testing methodology and a platform for such tests. Remote user testing is the future of usability tests, so introducing new methods and techniques to improve them is a very beneficial thing.

Acknowledgements. This work has been partially supported by the Polish Ministry of Science and Higher Education within the European Regional Development Fund, Grant No. POIG.01.03.01-00-008/08.

References

- [1] Andreasen, M.S., Nielsen, H.V., Schröder, S.O., Stage, J.: What Happened to Remote Usability Testing? An Empirical Study of Three Methods (20.03.2011), <http://www.takebay.net/data/chi07/docs/p1405.pdf>
- [2] Brush B., Ames M., Davis J.: A Comparison of Synchronous Remote and Local Usability Studies for an Expert Interface, <http://delivery.acm.org/10.1145/990000/986018/p1179-brush.pdf> (21.03.2001)
- [3] Duchowski, A.T.: Eye tracking methodology: Theory and practice, pp. 205–300. Springer-Verlag Ltd., London (2003)
- [4] Moha, N., Li, Q., Seffah, A., Michel, G.: Towards a Platform for Usability Remote Tests via Internet, http://www.ptidej.net/Members/mohanaou/paper/OZCHI2004/OZCHI2004_Moha.pdf (20.03.2011)
- [5] Mohamed, A.O., Pereira Da Silva, M., Courbolay, V.: A history of eye gaze tracking (2007), http://hal.archives-ou-vertes.fr/docs/00/21/59/67/PDF/Rapport_interne_1.pdf (12.03.2010)
- [6] Oztoprak, A., Erbug, C.: Field versus Laboratory Usability Testing: a First Comparison, http://www.aydinoztoprak.com/images/HFES_Oztoprak_.pdf (21.03.2011)
- [7] Petrie, H., Hamilton, F., King, N., Pavan, P.: Remote Usability Evaluations with Disabled People. In: CHI 2006 Proceedings, Montréal, Québec, Canada (2006), <http://www-course.cs.york.ac.uk/rmh/p1133-petrie.pdf> (20.03.2011)
- [8] Scholtz, J.: Adaptation of Traditional Usability Testing Methods for Remote Testing, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.2141&rep=rep1&type=pdf> (21.03.2011)
- [9] Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., Bergel, M.: An Empirical Comparison of Lab and Remote Usability Testing of Web Sites, <http://home.comcast.net/~tomtullis/publications/RemoteVsLab.pdf>

A Means-End Analysis of Consumers' Perceptions of Virtual World Affordances for E-commerce

Minh Quang Tran¹, Shailey Minocha¹, Dave Roberts¹,
Angus Laing², and Darren Langdridge³

¹Centre for Research in Computing, The Open University, UK

²School of Business and Economics, Loughborough University, UK

³Department of Psychology, The Open University, UK

{m.tran, s.minocha, d.roberts, d.langdridge}@open.ac.uk,
a.w.laing@lboro.ac.uk

Abstract. Virtual worlds are three-dimensional (3D) persistent multi-user online environments where users interact through avatars. The affordances of virtual worlds can be useful for business-to-consumer e-commerce. Moreover, affordances of virtual worlds can complement affordances of websites to provide consumers with an enhanced e-commerce experience. We investigated which affordances of virtual worlds can enhance consumers' experiences on e-commerce websites. We conducted laddering interviews with 30 virtual world consumers to understand their perceptions of virtual world affordances. A means-end analysis was then applied to the interview data. The results suggest co-presence, product discovery, 3D product experience, greater interactivity with products and sociability are some of the key virtual world affordances for consumers. We discuss theoretical implications of the research using dimensions from the Technology Acceptance Model. We also discuss practical implications, such as how virtual world affordances can be incorporated into the design of e-commerce websites.

Keywords: Consumer experience, e-commerce, interaction design, laddering interviews, means-end analysis, qualitative research, user experience, virtual worlds.

1 Introduction

Virtual worlds, such as Second Life [1], are being used by businesses for real world e-commerce. For example, Toyota and Reebok have created virtual stores and showrooms in Second Life for consumers to interact with virtual simulations of their real world products [2]. These and other real world businesses that create a presence in virtual worlds can increase their engagement with consumers [3] and enhance the value of their brands [4]. Given the potential benefits of using virtual worlds for real world e-commerce, researchers are beginning to investigate how virtual worlds can be better utilised by businesses [5].

In this paper, we investigated which virtual world affordances facilitate e-commerce from the consumer's perspective. For our study, we defined affordances as

features of a technology that determine ways a technology can be used [6]. Our investigation is based in Second Life. Second Life (<http://secondlife.com>) is a virtual world that was launched in 2003 and continues to attract 800,000 visitors each month. It supports many activities, including education, socialising, gaming and e-commerce. The e-commerce activity in Second Life mostly involves virtual items [7]. However, an understanding of the affordances for e-commerce with virtual items is useful for understanding how e-commerce in virtual worlds with real items can occur.

Some virtual world affordances used for e-commerce with virtual items are likely to be the same affordances used for e-commerce with real items. For example, virtual worlds afford the rendering of 3D objects in a 3D environment. This affordance allows consumers to have a virtual experience with real or virtual products, which can help consumers to learn about and evaluate products before buying [8]. Furthermore, the affordance of a persistent multi-user environment enables and sustains online communities. Online communities are known to influence consumers' experiences [9] regardless of whether the consumption is real or virtual. The aim of this paper is to identify the virtual world affordances that enhance consumers' real world e-commerce experiences. Our investigation has led to an understanding of when and why certain virtual world affordances are useful for e-commerce.

2 Research Question

The research question we have addressed is "*which affordances of virtual worlds can enhance consumers' e-commerce experiences?*" The research question has been investigated through laddering interviews and means-end analysis. These techniques are commonly applied in consumer research to understand consumers' preferences for brands, advertisements and products [10]. We have applied these techniques to understand consumers' perceptions of virtual world affordances for e-commerce. Laddering interviews were used to elicit the underlying reasons consumers have for preferring certain affordances. Means-end analysis was then used as a structured method to analyse the qualitative data from the laddering interviews [11].

3 Background

The theoretical framework of this study is based on the concept of the service encounter. The service encounter includes all the interactions that the consumer performs during the provision of a service [12]. In the context of e-commerce, the service encounter centres on the online purchase transaction. Moreover, the pre-purchase and post-purchase interactions are also a part of the service encounter. In this study, we have elicited consumers' perceptions of how virtual world affordances can enhance the e-commerce service encounter.

An important concept within the e-commerce service encounter is the consumer's ability to move between different online and offline service channels. Consumers choose appropriate service channels based on their needs during a service encounter. For example, additional support for the purchase can be provided through high street stores, telephone and mail-order catalogues [13]. This study starts with the premise

that virtual worlds are able to address certain needs that consumers have during the e-commerce service encounter.

Virtual worlds are said to incorporate the convenience of e-commerce websites with the richness of interactions of high street stores [14], but there is little empirical evidence to support this claim. Despite their potential, virtual worlds seem to be under-utilised for real world e-commerce. Currently, in April 2011, there are only a small number of virtual worlds that support online purchasing of real world products: some examples of virtual worlds designed specifically for e-commerce are Avaya (<http://avayalive.com>), NearWorld (<http://london.nearglobal.com>), and Trillenium (<http://www.trillenium.com>). Part of the motivation for this paper is to examine the reasons for the under-utilisation of virtual worlds for real world e-commerce and to understand what can be done to increase their usage.

According to the Technology Acceptance Model [15], the usage of technology is determined mainly by perceived usefulness and perceived ease-of-use. However, given the relative novelty of virtual worlds, our investigation was conducted in an exploratory manner. We did not limit the study to only investigate pre-determined dimensions of technology use, such as perceived usefulness and perceived ease-of-use; we allowed other dimensions to emerge from the data as well.

4 Data Collection

We conducted 30 interviews with participants who had e-commerce experience in Second Life. The participants also had e-commerce experience on websites. Participants were recruited and interviewed in Second Life through the use of avatars. The recruitment procedure and interview protocol were approved by our university's ethics committee. We also followed Second Life's community standards [16] to ensure that the study was not violating Second Life's usage policy.

To locate participants, we visited shopping and socialising areas in Second Life. In these areas, we scouted for prospective participants by first reading their user profiles. User profiles contain public information provided voluntarily by the user. The user profiles were utilised as a preliminary screening tool. Based on the profiles, we could identify experienced users of Second Life and we could identify users who had shopping experiences in Second Life. Prospective participants were then sent a private text message (similar to an instant message (IM)) inviting them to participate in the research study. Those that agreed were given a virtual note card containing detailed information about the study and our contact information. A second note card was given to participants as a consent form. Participants were then asked if they were over 18. If they were over 18, they were asked to consent to participate through text messaging.

Demographic information (real life age, real life gender, real life occupation, and country of residence for participants) was requested from participants at the end of each interview. Providing demographic information was optional. Thus, some participants did not reveal all their demographic information. With the information provided to us, we know that 13 males and 13 females participated. Four participants did not reveal their gender. Participants resided in USA (7), UK (3), Romania (3), Spain (2), New Zealand (2), Italy (2), India (2), France (1), Belgium (1), Australia (1) and Germany (1). Four participants did not reveal their country of residence. Participant's

ages ranged from 18 to 55. Eight participants did not reveal their age. The mean age of known participants was 29.1.

All the interviews were conducted in Second Life, using private text messaging. Interviews typically lasted 45 minutes to 1 hour. The interviews were conducted in English. Interviews were conducted between May 2010 and Dec 2010. Participants were given an honorarium worth two Great Britain pounds that was paid in Linden Dollars, Second Life's currency.

The interviews started with background questions to elicit participant's experiences of shopping in Second Life and shopping on websites. Participants were then asked to recall features of websites that they liked and disliked. This question was to prime participants so that they could more easily answer the main interview question. The main interview question was, "*can you think of Second Life features that would make shopping on websites better or easier?*"

As participants recalled features, we proceeded to apply the laddering technique. The laddering technique involved asking a series of probing questions, such as '*why is the feature important to you?*' The probes were meant to uncover participants' preferences for specific features of virtual worlds during the e-commerce service encounter.

5 Data Analysis

We analysed the interview data using means-end chain analysis [17]. The analysis started by extracting ladders, or means-end chains, from the data. Each ladder contains three components: an attribute, a consequence and a value.

- An *attribute* refers to any feature of the virtual world, but specifically something that can be interacted with or supports an interaction.
- A *consequence* is the outcome of performing the interaction.
- A *value* is the psychological need or desire associated with the consequence.

Each ladder represents participants' reasoning about why they think a feature of the virtual world is useful for e-commerce. An example of a ladder is: having avatars (*attribute*) to interact with real people (*consequence*) to feel a connection with others (*value*). Extracting ladders involved two steps: making annotations and linking the annotations.

1. Annotating involves identifying relevant segments of the data and describing them either as referring to an attribute, a consequence or a value. An additional description is given to the annotation that refers to its meaning. For example, an annotation could be: [*value: connection with others*]. For this annotation, '*value*' refers to the ladder component and '*connection with others*' refers to its meaning.
2. Linking annotations involves making explicit the connection between specific attributes, consequences and values that are mentioned by participants. The purpose is to create complete ladders with each of the three components: an attribute, a consequence and a value.

In the excerpt below, we show an example of how the data is annotated and linked to create a ladder. The three components for this ladder are: [attribute: *having avatars*][consequence: *interact with real people*][value: *connection with others*].

Interviewer: Can you think about how things you know about Second Life may help make shopping on a website better?

Consumer27: The sense of companionship [value: connection to others]. When I shop in [Second Life] it's usually with a couple of girlfriends [consequence: interact with real people].

Interviewer: And why is it important that you can shop with others?

Consumer27: I value their opinions on style and I like to share in their pleasure at finding something that looks good on them. It's a social activity.

Interviewer: Do you think there is something else about Second Life that helps with the companionship?

Consumer27: <nods> well the high-bandwidth communication

Interviewer: What do you mean by that?

Consumer27: Not just chat, but being able to share visually

Interviewer: Share visually?

Consumer27: mmHmm (sic) I can try something on and my friends tell me how it looks on me [...] the 3D models could be made off of real clothes using real photo textures so the ingredients are there for this kind of experience. [attribute: having avatars]

The data was annotated by using a bottom-up strategy, that is, by working from the data instead of using pre-determined concepts. Through several iterations, the annotations were eventually consolidated into a set of 51 codes, comprising of 16 attributes, 19 consequences and 16 values. We list the codes in the next section.

5.1 Codes Derived from the Annotation Process

Table 1 shows our codes and their frequency of appearance in the data. The percentage represents the frequency of the code relative to the total number of ladders.

Table 1. Codes for attributes, consequences and values

#	Code Name	Alternate Description	%
Attributes of Second Life			
1	Store attendants	Getting immediate help from business	10.5%
2	Teleportation	Instantly travelling to locations	0.9%
3	High graphic detail	High resolution of images and objects	2.6%
4	User generated content	Other users create objects in the world	1.8%
5	Access to websites	Easy to switch to and from websites	4.4%
6	Content creation tools	Tools for modifying in-world objects	0.9%
7	3D environment	Environment is rendered in 3D	15.8%
8	3D objects	Objects are rendered in 3D	21.1%
9	Multi-user	Sharing same environment with others	9.6%
10	Having avatars	Users embodied through an avatar	7.9%

Table 1. (Continued)

#	Code Name	Alternate Description	%
Attributes of Second Life			
11	Wide selection	Wide selection of products to browse	1.8%
12	Synchronous interactions	Interactions happen in real-time	4.4%
13	Alternative payment	Payment does not require a website	3.5%
14	Real world stores	Real stores have presence in-world	2.6%
15	Social network	Supports social networking	9.6%
16	Multimedia support	Support for audio and visual media	2.6%
Consequences of having the attributes			
17	Avoid inconvenience	Avoid travel and weather of real world	3.5%
18	Interact with real people	Real people are controlling the avatars	14.0%
19	Remain logged in	No need to exit virtual world	1.8%
20	Visually presented	Products are visually represented	11.4%
21	Interaction history	Saved history of interactions	2.6%
22	Can listen to music	Hear music and ambient sounds	0.9%
23	Interactions are in-world	Familiar navigation and interface	2.6%
24	Fast purchase transaction	Transactions are done in 'one click'	1.8%
25	Shop with friends	Can shop in friend's presence	4.4%
26	Maintain anonymity	Person remains anonymous	0.9%
27	Learn about products	Feel they know more about products	15.8%
28	Attracted to products	Product becomes desirable	3.5%
29	More reason to buy	Gain reasons to make purchase	1.8%
30	No pressure to buy	No hassle from sales person	1.8%
31	Increased trust	Confidence in business or product	10.5%
32	Discover products	Discover products unexpectedly	7.0%
33	See something new	Impressed by the novel graphics	1.8%
34	Alternate identity	Create an alter ego	1.8%
35	Play with products	Interact with products	12.3%
Personal values expressed by consumers			
36	Save money	Not wasting money on bad purchase	2.6%
37	Get immediate reply	Being attended to immediately	1.8%
38	More knowledge	Having more knowledge of products	22.8%
39	New experience	Experiencing something new	5.3%
40	Safety	Personal information kept private	6.1%
41	Familiar experience	Interacting in a 'natural' way	3.5%
42	Feel cared for	Sense of being attended to	6.1%
43	Informed decisions	Confidence in purchase decision	13.2%
44	Connection to brand	Relationship with brand	0.9%
45	Connection to virtual	Relationship with virtual community	3.5%
46	Feel welcomed	Feeling at ease in virtual world	0.9%
47	Disconnected from real	Forgetting about real world	0.9%
48	Enjoyment	Experience is 'fun'	23.7%
49	Connection to others	Relationship with individuals	3.5%
50	Freedom to play	Allowed to experiment with products	2.6%
51	Save time	Get things done faster in virtual world	2.6%

The codes in Table 1 were used to derive ladders during the linking process.

5.2 Ladders Derived from the Linking Process

Alongside the annotation process, ladders were derived through the linking process described above. 114 ladders were derived from the data: a mean of 3.8 ladders per participant. The ladders were then entered into LadderUX. LadderUX is a data analysis software package for laddering research [18]. The software creates hierarchical value maps based on the formula described in Reynolds and Gutman's seminal paper on laddering [17]. First, a matrix is created that cross-links all the codes. Each cross-link between codes is given a score depending on how frequently a link between the codes appears. Then, a map is generated based on the scores. Only those scores which are above a defined cut-off level are shown in the map. The researcher defines the cut-off level. Cut-off levels from 3 to 5 are most common in laddering research. The cut-off level we have used for our map is 4 (see Figure 1, section 6.2).

6 Results

In this section, we summarise the results of the analysis. First, we discuss the attributes, consequences and values. Then, we discuss the hierarchical value map and finally, we re-visit the research question of this paper. For the results section, participants are referred to as consumers.

6.1 Summary of Attributes, Consequences and Values

Attributes

Attributes are features of virtual worlds. The most frequently mentioned attributes by consumers are shown in Table 2. Only attributes with a frequency higher than 9% are discussed in this paper. At the 9% cut-off, we found a close fit between the number of elements in the summary tables (e.g. Table 2) and the number of elements in the hierarchical value map. Attributes that are related are grouped and discussed together.

Table 2. Summary of most frequently mentioned attributes

Attribute Code	Code Frequency	Attribute Group	Grouped frequency
3D objects	21.1%	3D aspects	36.9%
3D environment	15.8%		
Store attendants	10.5%	Social aspects	29.7%
Social network	9.6%		
Multi-user	9.6%		

Attribute group: 3D aspects

The first grouping that emerged is 3D aspects, which was present in 36.9% of the ladders. This relates to 3D objects and the 3D environment of virtual worlds. It is not surprising these attributes are mentioned most frequently. The difference between the

3D environment in virtual worlds and the 2D display on websites is easy for consumers to notice.

Consumer3: *The layout when you shop in [Second Life] is way nicer than on the average website. I mean, ok, [Second Life] is 3D whereas a site is merely 2D. It's just that the presentation and the [angle of] perception are so much better used in [Second Life], which makes you wanna (sic) buy stuff.*

When consumers discuss the 3D aspects, they usually refer to navigation and perception. Navigation relates to being able to walk through a virtual store to browse virtual shelves and see rows of virtual products. This simulates the navigation experience of a real store. Perception is about having objects represented with depth (hence, 3D). Both navigation in a 3D environment and 3D perception add to the sense of realism.

Consumer16: *With 3D it gives it a little better feel of what you are getting into.*

Interviewer: *Can you say a bit more about what you mean by 'feel'?*

Consumer16: *Well to simplify. Compare Sega to Playstation 3. Sega is flat images, 16-bit. Play station has 3D graphics. Which appeals to people more? Playstation 3 of course because it gives people that sense of realism. It's as if the objects are there in front of you vs on paper.*

Consumers report the 3D aspects in terms of its visual appeal. "*The [3D] interface is more seductive than flat corporate web sites*" (Consumer5). Beyond visual appeal, 3D aspects allow greater interactivity.

Consumer23: *It's just fun [...] you feel like its [real life]*

Interviewer: *Can you tell me what is it about the '[real life] feeling' that makes it fun?*

Consumer23: *You just feel like you're actually in a shop, not flicking thru a mag (sic).*

Attribute group: Social aspects

The second grouping that emerged is social aspects, which was present in 29.7% of the ladders. Social aspects entail having other people around who are also participating in the shopping experience. These could be strangers, friends or sales personnel.

Consumer1: *Make [shopping] more attractive. Event organisation, product presentations, and people we can talk to about their products, etc. I am reminded of the Reebok SL shop. It was empty, strangely dead. Too bad, they were selling interesting things for a good price*

Interviewer: *Can you tell me about it being empty? Why is this a bad thing?*

Consumer1: *Their shop was like [16 km²] for only 10 articles and nobody in there. [...] It's like entering a huge empty mall. It makes me think about "night of the living dead".*

Consumers appreciate having others around, especially people they are familiar with and can talk to.

Consumer26: *When you go shopping it's nice especially with friends.*

Interviewer: *Why is it nice to shop with friends?*

Consumer26: *That it's not boring [...] you [can] talk and not walk alone through the store.*

Interaction with sales personnel is also frequently mentioned as an advantage of shopping in virtual worlds.

Consumer13: *[Store attendants can] help the customer find the product that best suits them, and answer questions about products. I guess going with the jewellery thing, I'm thinking of a boutique kind of experience.*

Consequences

Now we look at the consequences of utilising virtual worlds for e-commerce. Consequences relate to outcomes of utilising virtual world features. Table 3 lists consumer's most frequently mentioned consequences. As with the attributes, we only summarise the consequences with a frequency over 9%.

Table 3. Summary of most frequently mentioned consequences

Consequence Code	Code Frequency	Consequence Group	Grouped Frequency
Learn about products	15.8%	Product fidelity	39.5%
Play with products	12.3%		
Visually presented	11.4%		
Interact with real people	14.0%	Interact with real people	14.0%
Increased trust	10.5%	Increased trust	10.5%

Consequence group: Product fidelity

Product fidelity is a grouping of three consequences: learning about products, playing with products and having products visually presented. As a group, it was included in 39.5% of the ladders mentioned by consumers. Products that are displayed in virtual worlds can simulate more of a product's functions and provide a better view of its physical dimensions. Ultimately, product fidelity means consumers can get a better idea of what the product is actually like. This is related to the concept of virtual experience from Li et al. [8] who conclude that interacting with 3D models of products is closer to a "direct" experience with a product. Consequently, interacting with 3D models helps the consumer learn about products more effectively compared to learning from text and 2D images [19].

Consumer5: *You would be able to see the scale of things, if [the dimensions] were accurate. I could even imagine trying a bed out in [...] my house to see it looks good next to all of my other things.*

Consequence group: Interact with real people

Virtual worlds are persistent real-time communication platforms. These attributes allow for real-time interactions with other people. Being able to interact with others is mentioned in 14.0% of the ladders. Interactions can be as simple as text chats with sales support personnel, friends or other consumers. However, interactions can also be more advanced; such as witnessing 'live' musical performances.

Interviewer: Is this different from how [musicians] promote themselves on a website?

Consumer28: YES (sic) because they can stream in and perform live

Interviewer: So is there something about a 'live' performance that makes it different?

Consumer28: It's an immersive experience that really feels like an intimate performance. Music is always best live when the musician can relate to the audience. Just like in [real life] the performer can connect with the audience.

Consequence group: Increased trust

Increased trust is about gaining confidence to make a purchase decision or gaining confidence to deal with a business. Consumers mention increased trust in 10.5% of the ladders. Trust is mentioned in relation to certain features of virtual worlds. For example, trust was mentioned alongside interacting with real people and interacting with high-fidelity products. It is also contingent on the type of products or services being bought.

Consumer6: I trust in many [businesses]. [...] For example I know you can pay for English class in [Second Life]. I think I could trust [them].

Values

Values relate to psychological needs and motivations. In Table 4, we summarise most frequently mentioned values by consumers.

Table 4. Summary of frequently mentioned values

Value Code	Code frequency	Value Group	Grouped frequency
More knowledge	22.8%	Knowledge	36.0%
Informed decisions	13.2%		
Enjoyment	23.7%	Enjoyment	23.7%

Value group: Knowledge

Knowledge implies the need to know more about products or the need to make informed decisions about buying products. It is the value in 36.0% of the ladders. Gaining knowledge is important because consumers ultimately have to make a purchase

decision. Understandably, consumers want to gather as much information as they can about products (or the business) so that they make the right decision.

Consumer19: Maybe you don't know how it looks like; you could get an idea here [in Second Life].

Consumer21: You don't want to spend money on things you will never use or looks bad.

Value group: Enjoyment

Enjoyment relates to having a fun and stress-free experience. Many consumers think that shopping in virtual worlds can be more enjoyable than shopping on websites: 23.7% of the ladders have enjoyment as their value.

Consumer9: You can go anywhere [in Second Life] at the click of a button. Shop anywhere all round the world, buy ANYTHING :D (sic). No limits.

Interviewer: And you enjoy this?

Consumer9: yup

Interviewer: What makes being able to [go] anywhere enjoyable?

Consumer9: It's free, fast, can go with anyone

Interviewer: Do you get the same feeling when you browse websites?

Consumer9: ummm naah (sic). That gets boring lol

Interviewer: Why is going from website to website boring, but going from shop to shop in [Second Life] fun?

Consumer9: Because you can walk around here, be with others, try things on, meet others.

6.2 Relationship between Attributes, Consequences and Values

Figure 1 shows the hierarchal value map based on our analysis (please see section 5.2). The thickness of the lines between codes in the map indicates a higher link frequency between the codes. Functional consequences noted in the map relate to observable outcomes, whereas psycho-social consequences relate to personal outcomes (such as a change in emotion or thought).

From the map, we can derive the reasons *why* and *when* consumers may choose virtual worlds over websites. We call these reasons “motivational patterns”, a term from Gengler et al. [20]. To derive motivational patterns, we start with the values at the top of the map and then follow the links downwards.

- Motivational pattern 1 is about enjoying the experience. This comes often from discovering new products and interacting with people. The 3D multi-user environment facilitates discovering new products and interacting with people.
- Motivational pattern 2 is about acquiring information about products. Consumers gather information about products by interacting with 3D simulations of products in virtual stores. Consumers also learn by interacting with store attendants in 3D spaces.

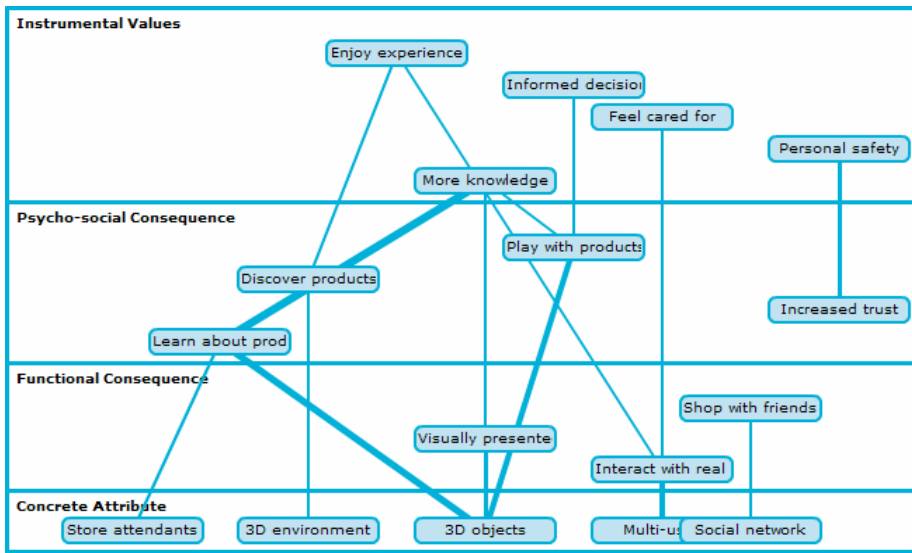


Fig. 1. Hierarchical value map with cut-off level at 4

- Motivational pattern 3 is about making informed decisions. This is possible because consumers can interact with virtual simulations of products, which are rendered and displayed in a 3D environment.
- Motivational pattern 4 is about feeling cared for. Consumers feel cared for because they are interacting with real people, which is possible because a virtual world is a multi-user environment.
- Motivational pattern 5 is about personal safety. Personal safety refers to concerns about privacy: for example, protecting personal information and banking details. Personal safety and privacy are related to trust. Consumers perceive interactions in virtual worlds to be more trustworthy because the interactions are interpersonal and because there are ways to verify a business's real world identity from websites or their stores.

These motivational patterns provide an understanding of how virtual worlds can enhance the e-commerce experience from the consumer's perspective. In the next subsection, we discuss affordances and answer the original research question.

6.3 Which Affordances of Virtual Worlds Can Enhance Consumers' E-commerce Experiences?

The research question for this study was, "*which affordances of virtual worlds can enhance consumers' e-commerce experiences?*" We answer this question by returning to the hierarchical value map (Figure 1, section 6.2). To identify the affordances, we start with the attributes from the map and then follow the links from the attributes upwards. Thus, the affordances that can enhance consumers' experiences are:

- Affordance 1: Multi-user environments to allow shopping with others
- Affordance 2: 3D environments to allow products to be easily discovered
- Affordance 3: Store attendants and 3D objects to support product learning
- Affordance 4: 3D objects to support more interactivity with products
- Affordance 5: Social networks to increase consumer's trust

We discuss implications of these affordances and consumers' perceptions of the affordances in the next section.

7 Implications for Theory

7.1 Technology Acceptance

Understanding consumers' perceptions of virtual world affordances can help predict why virtual worlds may be adopted by consumers for the e-commerce service encounter. Two dimensions of perception that are highly correlated with technology acceptance are perceived usefulness and perceived ease-of-use [21]. Usefulness refers to how proficiently a user will be able to complete a task with a technology. Ease-of-use refers to how easily a user can learn and utilise a technology. Both dimensions will be discussed in relation to virtual world affordances.

This study suggests that virtual worlds are perceived to be useful when the task is to learn about products. Virtual worlds are also perceived to be useful for interacting with other people. Therefore, consumers are likely to use virtual worlds for e-commerce when they want to learn about products or interact with others in a more effective and efficient manner.

This study indirectly addresses the dimension of perceived ease-of-use. Perceived ease-of-use is alluded to by our findings on enjoyment. Enjoyment, more commonly conceptualised as satisfaction, is related to usability. In turn, usability is related to ease-of-use [22]. Consumers may choose to use virtual worlds for e-commerce when they want to enjoy the e-commerce experience. To what extent the hedonic aspect (enjoyment) and utilitarian aspects (efficiency and effectiveness) overlap is uncertain. However, our study suggests that consumers perceive virtual worlds have the ability to offer both enjoyment and utility.

Four other dimensions of consumer's perceptions also emerged from our study. The dimensions are: compatibility, perceived service quality, perceived trust and information richness. These dimensions are also part of a website technology acceptance model by Chen and Tan [23].

Compatibility refers to the overlap between the consumers' personal beliefs and what they believe the technology represents. In other words, it relates to prejudices consumers' hold for or against a technology. The concept of compatibility originates from Rogers' theory on diffusion of innovations [24]. Bessiere et al. [25] describes an example of how compatibility can influence virtual world acceptance. They found users in their study associated virtual worlds "primarily with socialising and playing, rather than working". The belief that virtual worlds offer a playful rather than a serious environment was an adoption barrier. Similarly, consumers may have prejudices against adopting virtual worlds for e-commerce as they may think that virtual worlds are not for making real world transactions.

Service quality is another dimension of consumers' perception. Service quality refers to the expectations of consumers about the service they will receive from a business. One facet of service quality is empathy, which is defined as "caring, individualised attention" towards consumers [26]. Our study found that empathy could be an advantage of shopping in virtual worlds. Consumers in our study explained that virtual worlds allow them to interact synchronously with sales personnel; the sales personnel were then able to answer questions immediately and in a personalised manner. Therefore, the interactions in virtual worlds were perceived as involving more empathy.

Consumers in our study were perceived trust as another advantage of virtual worlds. However, trust varies depending on the situation. Consumers were not trusting of virtual worlds when making a transaction, but they were trusting of virtual worlds for advice. Consumers trust advice in virtual worlds because it comes from real people with whom they feel were co-present. The realism of the shopping environment may also be a reason for consumer's increased trust [27].

Information richness is the last dimension we discuss. Information richness is defined as "quality of product information and the extent of product comparison" [23]. Information richness enables the consumer to make an informed purchase decision. It relates to product learning, but also to the search, sense-making and decision-making process. Several consumers in our study discussed aspects relating to information richness. The 3D environment and 3D simulations help increase information richness. The interactions with other consumers and sales personnel also help consumers gather more information while in virtual worlds. Therefore, in terms of technology acceptance, information richness is another explanation why consumers' may choose virtual worlds for e-commerce.

7.2 Multi-channel Consumption and the Service Encounter

In the previous sub-section, we discussed some opportunities and barriers for adoption of virtual worlds for e-commerce. In this sub-section, we start with the assumption that virtual worlds are already a viable option for e-commerce. Another question then arises: when will consumers choose to use virtual worlds instead of other service channels?

Consumers have many options of channels during a service encounter, such as websites, high street, telephone, and so on. Virtual worlds are not likely to replace any of the existing service channels; more likely, virtual worlds will become another option to complement the existing channels. Research on multi-channel consumption behaviour suggests perceived risk, need for interpersonal contact, convenience, and price differential are factors that determine which channel consumers will choose for the service encounter [28]. Furthermore, research suggests consumer's choice of channel is context-specific and idiosyncratic [13]. Our study identifies some contexts where virtual worlds can enhance the service encounter, such as when consumers need to talk to sales personnel, have the desire to browse or when they would like to interact with 3D simulations of products.

Different service channels have different affordances. Our study contributes towards identifying the affordances of virtual worlds that make virtual worlds a unique service channel. Understanding the key affordances of virtual worlds can help designers make decisions on how to best utilise virtual worlds during a service encounter.

For example, the consumer may want to learn about a product without leaving their home, but is disappointed with the images or product descriptions on websites. The consumer could instead enter a virtual world to interact with 3D simulations and discuss the product with sales personnel, or ask family and friends to accompany them to 3D stores to help in the decision-making process. Then, the consumer could return to the website to complete the purchase transactions because they may perceive the website is trustworthy and the payment will be handled securely.

8 Practical Implications for E-commerce on Websites

In some cases, it may not be practical to have the service encounter entirely in a virtual world. In these cases, affordances can be abstracted and incorporated into the design of e-commerce websites. In this section, we discuss five key affordances of virtual worlds for e-commerce that can be incorporated into websites. The five affordances we discuss are: facilitating co-presence, greater support for product discovery, virtual experience, playful experience with products, and building trust through interpersonal interactions. These affordances are related to the affordances mentioned in section 6.3.

Allowing the feeling of co-presence

Websites should build features to support co-presence [29]. For example, co-presence can be supported through a chat feature embedded into websites. Shopping on websites would then become a collaborative activity, allowing consumers to shop ‘together’. There is an inherent trade-off in privacy and anonymity with this affordance. However, in some situations, consumers may desire it. Based on our study, this affordance is related to the personal desire to enjoy the experience.

Greater product discovery and browsing opportunities

Websites should build features for product discovery through browsing. Browsing is a different kind of information-seeking behaviour from searching [30]. With browsing, the consumer does not necessarily know what they are looking for. Many e-commerce websites are designed for efficient search, but not for efficient browsing. Our study suggests that consumers appreciate the ability to browse as well during e-commerce. 3D environments, which have rich visuals, can facilitate browsing. On websites, rich visuals can be combined with intelligent recommender systems to provide consumers opportunities to browse products online.

Having virtual experiences with products

Websites should include features to provide an increased interactive experience with products to reproduce the virtual experiences that consumers have from using virtual worlds. Virtual experiences results from interacting with 3D rendered products that are “moveable, rotatable, zoomable, customizable, and animated” [19]. Another advantage of incorporating virtual experiences is that consumers feel more engaged

when they learn with 3D objects compared to learning about products through text and 2D images [8].

Having playful experiences with products

Websites should build features to allow consumers to play with products. This is closely related to the affordance of having virtual experiences. However, playful experiences result in more than just learning about a product. Playful experiences result in engagement and enjoyment as well. The key feature related to playfulness is flexibility to manipulate virtual products in different contexts. For example, consumers discussed mixing and matching clothing items or rearranging different furniture pieces in a room. To support play and playfulness, websites should consider increasing their product selection (variety), adding elements onto their website that are modifiable and adding elements that do not directly serve utilitarian purposes [31].

Building trust through social networks and interpersonal interaction

Websites should support some form of social networking to allow consumers to connect and communicate with each other. This affordance is perceived to enhance trust. Trust, in this case, might come from having social capital. This is sometimes called "social trust" and it is correlated with a higher probability that consumers will make purchase transactions [32]. Large online retailers (e.g. amazon.com and ebay.com) have incorporated this feature with success by allowing customers to review products on their websites.

9 Limitations and Directions for Further Study

This study provides an empirically-grounded understanding of virtual world affordances for e-commerce. However, as with any empirical research, validity and generalisability should be considered in relation to how the data was collected and sampled. The participant sample for our study was experienced users of virtual worlds. Therefore, we presume the participants already enjoy using virtual worlds. They may also be more trusting of interactions in virtual worlds because of their familiarity with places, people and the interface in virtual worlds. Participants who are inexperienced with virtual worlds may have different perceptions compared to the participants in our study. Further studies with participants who are inexperienced with virtual worlds could provide more perspectives to complement the findings of this study.

Another point worth considering is that the experiences of our participants were based on their consumption of virtual items. Although the consumption processes may be similar for real and virtual items, there is still uncertainty about how much can be generalised about the use of the virtual world affordances for purchase of real world items. Further studies may want to look at consumption tasks in virtual worlds involving real items. However, this latter option is difficult given the scarcity of real world e-commerce in virtual worlds. Another option would be to perform laboratory studies where the consumption tasks and the choice of products can be controlled.

The communication mode used to collect data may be another limitation on the data. Using private text messages could influence the type and range of responses

elicited during the laddering interviews. Other methods of collecting qualitative data, such as face-to-face interviews or online voice interviews, may provide a different range of responses. For example, if interviewed face-to-face, participants may be more aware of the benefits of interpersonal contact.

Finally, consumers' perceptions were based on interactions in Second Life. Second Life is one amongst many virtual worlds. The service encounter in Second Life is likely to be different from the service encounters in other virtual worlds because there would be differences between storefronts, personnel and interface designs. One solution would be to replicate the study using other virtual worlds, even with those that are not designed specifically for e-commerce, such as World of Warcraft [33].

10 Conclusion

This study explored the potential for utilising virtual worlds as an e-commerce service channel. The study shows that consumers' perceive virtual worlds to retain the convenience of an online service channel, but have additional affordances normally associated with offline service channels. These additional affordances are co-presence, product discovery, virtual experience with products, greater freedom to interact with products, and sociability. By identifying these affordances, and identifying consumers' perceptions related to them, we have provided some direction for future research. We also provided priorities for interaction design of business-to-consumer e-commerce on websites and in virtual worlds.

References

1. Rymaszewski, M., Au, W.J., Ondrejka, C., Platel, R., Gorden, S.V., Cezanne, J., et al.: *Second Life: The Official Guide*, 2nd edn., pp. 360–362. Wiley Publishing, Indiana (2008)
2. Goel, L., Prokopec, S.: If You Build it Will They Come? An Empirical Investigation of Consumer Perceptions and Strategy in Virtual Worlds. *Electron. Commer. Res.* 9, 115–134 (2009)
3. Park, S.R., Nah, F.F., Dewester, D., Eschenbrenner, B.: Virtual World Affordances: Enhancing Brand Value By Virtual World Affordances: Enhancing Brand Value. *Journal of Virtual Worlds Research* 1(2), 1–18 (2008)
4. Barnes, S., Mattsson, J.: Brand Value in Virtual Worlds: An Axiological Approach. *Journal of Electronic Commerce Research* 9(3), 195–206 (2008)
5. Cagnina, M., Poian, M.: Beyond E-Business Models: The Road to Virtual Worlds. *Electron. Commerce Res.* 9(1), 49–75 (2009)
6. Norman, D.A.: Affordance, Conventions, and Design. *Interactions*, 38–42 (May/June 1999)
7. Bakshy, E., Simmons, M.P., Huffaker, D.A., Teng, C.Y., Adamic, L.A.: The Social Dynamics of Economic Activity in a Virtual World. In: *Fourth International AAAI Conference on Weblogs and Social Media* (2010)
8. Li, H., Daugherty, T., Biocca, F.: Characteristics of Virtual Experience in Electronic Commerce: A Protocol Analysis. *J. Interact. Market.* 15(3), 13–30 (2001)
9. Kozinets, R.V.: E-Tribalised marketing?: The Strategic Implications of Virtual Communities of Consumption. *Eur. Manag. J.* 17(3), 252–264 (1999)

10. Wansink, B.: Using Laddering to Understand and Leverage a Brand's Equity. *Qual. Market. Res. Int. J.* 6(2), 111–118 (2003)
11. Veludo-De-Oliveira, T.M., Ikeda, A.A.: Discussing Laddering Application by the Means-End Chain Theory. *The Qualitative Report* 11(4), 626–642 (2006)
12. Bitner, M.J., Booms, B.H., Tetreault, M.S.: The Service Encounter: Diagnosing Favorable and Unfavorable Incidents. *J. Market.* 54(1), 71–84 (1990)
13. Dijk, G.V., Minocha, S., Laing, A.: Consumers, Channels and Communication: Online and Online Communication in Service Consumption. *Interact. Comput.* 19, 7–19 (2007)
14. Hemp, P.: Are You Ready for E-tailing 2.0? *Harvard Business Review*, 28 (June 2006)
15. Davis, F.D.: Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13(3), 391 (1989)
16. Second Life Community Standards (2011), <http://secondlife.com/corporate/cs.php>
17. Reynolds, T.J., Gutman, J.: Laddering Theory, Method, Analysis, and Interpretation. *J. Advert. Res.* 28(1), 11–32 (1988)
18. Abeeel, V., Zaman, B.: Laddering the User Experience, http://ladderux.org/papers/Laddering_the_User_Experience.pdf
19. Zhenhui, J., Benbasat, I.: Virtual Product Experience: Effects of Visual and Functional Control of Products on Perceived Diagnosticity and Flow in Electronic Shopping. *J. Manag. Inform. Syst.* 21(3), 111–147 (2004)
20. Gengler, C., Mulvey, M., Oglethorpe, J.: A Means-End Analysis of Mothers' Infant Feeding Choices. *J. Publ. Pol.* 18(2), 172–188 (1999)
21. Legris, P., Ingham, J., Collette, P.: Why Do People Use Information Technology? A Critical Review of the Technology Acceptance Model. *Inform. Manag.* 40(3), 191–204 (2003)
22. Lederer, A.L., Maupin, D.J., Sena, M.P., Zhuang, Y.: The Technology Acceptance Model and the World Wide Web. *Decis. Support Syst.* 29(3), 269–282 (2000)
23. Chen, L.D., Tan, J.: Technology Adaptation in E-commerce: Key Determinants of Virtual Store Acceptance. *Eur. Manag. J.* 22(1), 74–86 (2004)
24. Rogers, E.M.: *Diffusion of Innovations*, 5th edn. The Free Press, New York (2003)
25. Bessiere, K., Ellis, J.B., Kellogg, W.A.: Acquiring a Professional “Second Life:” Problems and Prospects for the Use of Virtual Worlds in Business. In: 27th International Conference on Human Factors in Computing Systems, pp. 2883–2898 (2009)
26. Buttle, F.: SERVQUAL: Review, Critique, Research Agenda. *Eur. J. Market.* 30(1), 8–32 (1996)
27. Nassiri, N.: Increasing Trust Through the Use of 3D E-commerce Environment. In: 23rd ACM Symposium on Applied Computing, pp. 1463–1466 (2008)
28. Black, N.J., Lockett, A., Ennew, C., Winklhofer, H., McKechnie, S.: Modelling Consumer Choice of Distribution Channels: An Illustration from Financial Services. *Int. J. Bank Market.* 20(4), 161–173 (2002)
29. Gerhard, M., Moore, D., Hobbs, D.: Embodiment and Copresence in Collaborative Interfaces. *Int. J. Hum. Comput. Interact.* 61, 453–480 (2004)
30. Toms, E.: Understanding and Facilitating the Browsing of Electronic Text. *International Journal of Human-Computer Studies* 52(3), 423–452 (2000)
31. Chung, J., Tan, F.B.: Antecedents of Perceived Playfulness: an Exploratory Study on User Acceptance of General Information-Searching Websites. *Inform. Manag.* 41, 869–881 (2004)
32. Mutz, D.C.: Social Trust and E-Commerce: Experimental Evidence for the Effects of Social Trust on Individuals' Economic Behavior. *Publ. Opin. Q* 69(3), 393–416 (2005)
33. Bainbridge, W.S.: The Scientific Research Potential of Virtual Worlds. *Science* 317 (2007)

Improving Users' Consistency When Recalling Location Sharing Preferences

Jayant Venkatanathan¹, Denzil Ferreira¹, Michael Benisch², Jiali Lin², Evangelos Karapanos¹, Vassilis Kostakos¹, Norman Sadeh², and Eran Toch²

¹ Madeira Interactive Technologies Institute, University of Madeira

² School of Computer Science, Carnegie Mellon University

{vjayant, denzil.ferreira, e.karapanos, vk}@m-iti.org

{mbenisch, jialiul, sadeh, eran}@cs.cmu.edu

Abstract. This paper presents a study of the effect of one instance of contextual cues, trajectory reminders, on the recollection of location sharing preferences elicited using a retrospective protocol. Trajectory reminders are user interface elements that indicate for a particular location of a person's trail across a city the locations visited before and after. The results of the study show that reminding users where they have been before and after a specific visited location can elicit more consistent responses in terms of stated location sharing preferences for that location visit. This paper argues that trajectory reminders are useful when collecting preference data with retrospective protocols because they can improve the quality of the collected data.

Keywords: Location sharing preferences, consistency, retrospective protocols.

1 Introduction

Location sharing applications are gaining wide adoption, with a number of commercial systems now available on the market, including Foursquare and Facebook Places. Such services are frequently used in the context of online social networks, whereby one's real-time location becomes yet another sharable aspect of one's online profile. With the increasing adoption of online location sharing services, understanding users' preferences and needs in terms of location sharing becomes crucial. Due to the nature of these services, it is methodologically difficult to elicit users' preferences in real time since this is likely to interrupt users' ongoing activities. Hence retrospection is a crucial methodological tool for eliciting users' preferences in this domain.

However, retrospective methods are susceptible to producing unreliable results. Due to their situated nature, location sharing preferences may depend on multiple contextual variables. Retrospective protocols may not reliably capture these characteristics and therefore elicit unreliable responses from users.

This paper presents a study that assesses the test-retest reliability of retrospective protocols, and introduces a technique for increasing the reliability of elicited responses in such protocols when collecting location sharing preferences. Grounded

on experience-reconstruction theory [9], the technique entails the introduction of trajectory reminders in the data-collection GUI to help users recall episodic information that may be used to infer one's preferences for location sharing.

2 Related Work

There is an increasing amount of work on understanding users' location-privacy needs in ubiquitous and location-aware systems relying on techniques such as diary studies [1], interviews [5], surveys [8], scenarios [12] and lab and field observations [2]. A significant decision in understanding users' privacy needs in relation to location sharing is the methodology by which privacy preferences are elicited.

In attempting to elicit privacy preferences regarding location sharing, one could ask participants to provide an overall estimate of their preferences for a given location, such as one's workplace. However, these often-called global measures have been shown to underestimate the variability in perceptions and preferences as people often fail or incompletely reconstruct the particular context of each situation [10]. Robinson & Clore [9] proposed a four-stage accessibility model of experiential information. At the heart of their model lies the distinction between episodic and semantic memory [13]. While episodic memory "is specific to a particular event from the past, semantic memory is not tied to any particular event but rather consists of certain generalizations (i.e. beliefs) that are rarely updated" [9]. In reconstructing one's emotions during an event, Robinson & Clore's [9] model argues that he or she first attempts to recall contextual cues from episodic memory. When episodic memories become inaccessible (for instance because the event is located further in the past), people will shift to semantic memory.

The Experience Sampling Method (ESM) [4] attempts to avoid such retrospection and rationalization biases through probing the participant to report on ongoing behaviors and experiences. One of the drawbacks of ESM, however, is its labor-intensive nature as it requires participants to interrupt their activities at numerous times within a day, while it may also miss important information when participants are not able to respond [15]. An alternative approach is the Day Reconstruction Method [6], a survey method that asks participants to recall in forward chronological order all experiences that took place in the previous day. Each experience is thus reconstructed within a temporal context of preceding and following ones. This is expected to cue more contextual information from episodic memory, and consequently recall the experienced emotion in a more valid and reliable manner. Kahneman et al. [6] showed that this method provides a surprisingly good approximation to Experience Sampling data, while providing the benefits of a retrospective method. Similarly Karapanos et al. [14] found that imposing a chronological order in the reconstruction of events resulted in an increase in the amount, the richness, and the test-retest reliability of recalled information.

In summary, literature suggests that if ESM is too intrusive then contextual cues should be used to help users reconstruct experiences from episodic memory. It follows that in asking participants to provide privacy-related preferences, a method that provides participants with contextual cues would elicit more reliable responses. To test this assumption in the context of location sharing privacy preferences, the

study described next compares the reliability of privacy preferences when those are elicited with vs. without the help of contextual cues, which in this case are instantiated as trajectory hints. Specifically, the tested assumption is that reminding a participant where they have been before and after a visited location will cue more contextual information from the experience being measured, therefore resulting in more reliable and consistent recall of privacy preferences.

3 Study

A study was designed to test the hypothesis: *Location Sharing privacy preferences elicited with the help of trajectory reminders will be more consistent than those elicited without the help of trajectory reminders.*

A total of 20 participants were recruited with an average age of 28 (max = 44, min = 21, median = 27, s.d = 5.1) , through announcements on email lists, online forums, and fliers distributed across the campuses of the University of Madeira (Portugal) and University of Oulu (Finland). No reward was offered to participants. 9 participants (all male) were allocated to System A (no trajectory reminders) while 11 (8 male) were allocated to System B (with trajectory reminders). This difference in numbers in participants and genders across the two conditions was due to dropouts.

Each participant was given an Android smart-phone equipped with GPS logging software and was instructed to use this phone as their primary phone to ensure that they kept it with them at all times. During registration participants were also asked to list the names of five people from each of their family, close friends and colleagues. Each participant was asked to use the phone for a period of 4 days, spanning both weekdays and weekends. During this period the phone recorded each participant's location. The software interface allowed the participants the option to temporarily disable the logging software should they wish to do so. Participants were instructed to upload their location history data at the end of each day after which they were required to immediately do an online questionnaire task.

When the location history was uploaded, the set of "important locations" was selected from the uploaded data. Specifically, the chosen locations were those where the participants spent at least 5 minutes within a 50 meter radius. Subsequently, participants were taken through a series of questionnaire pages, where each page displayed an important location on the map along with the details of when the participant was there. Depending on the system to which participants were allocated, the location was displayed on a map with or without trajectory reminders. These reminders were arrows that indicated where the participant had been before and after visiting a location (Figure 1). All locations were displayed in chronological order.

As each important location was shown to participants, they were asked to recall how much information they would have liked to share about that specific location at that specific time with one specific person (chosen at random) from each of the 4 different recipient groups . The information was entered in a scale (1 to 5) with the following sharing options: (1) do not share, (2) region, (3) city, (4) neighborhood, (5) exact address. Besides entering their location sharing preferences, participants had the option to indicate that a location was completely wrong or invalid, although this option was never chosen.

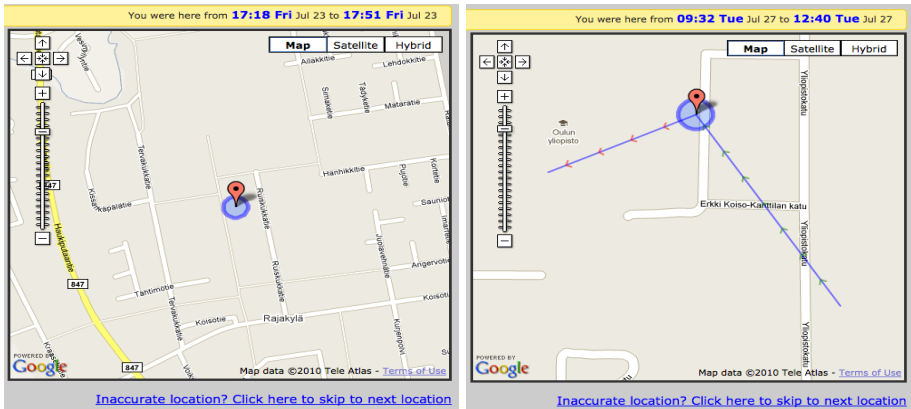


Fig. 1. The two versions of the map shown to participants. On the left is system A that shows no trajectory reminders. On the right is system B that shows trajectory reminders, i.e. the locations visited before and after the location in question.

A week after returning the phones, participants were asked to access the online system and re-enter their preferences for each location they visited while they were carrying the phone. At this stage participants could not see their earlier responses. The previously recorded locations were shown with the same details and in the same order as the first time participants gave their preferences. Participants gave their preferences for all of their important locations in a single session in order to minimize the strain and reduce drop-out rates.

4 Results

The study was conducted between May and September 2010. A total of 441 distinct location visits were recorded, with each visit lasting on average 110 minutes. For each distinct visit to a location the following data was recorded: time and duration of visit, privacy preference (on a scale 1-5) for each of 4 possible recipient groups (family, close friends, colleagues, strangers), and the system being used (A or B). Of these, the independent variables were “System” and “Recipient_group”. In addition, the difference in privacy preference was calculated by comparing the results from the two sets of questionnaires (the first was issued on the day of the visit to the location, the second was issued one week after the end of the study).

A chi-square test showed a significant effect of trajectory reminders ($X^2 = 43.5653$, $df = 8$, $p < 0.001$) and recipient type ($X^2 = 116.038$, $df = 24$, $p < 0.001$) on the variation of privacy preferences. In terms of magnitude of the effect on consistency, the mean absolute difference in consistency for system A (no trajectory reminders) is 0.16, and for system B (trajectory reminders) is 0.06.

Figure 2 shows the mean variation in participants' responses, grouped by recipient type and system type. In these results a positive variation means that the follow-up response was more liberal (i.e. shared more information), a negative variation means that the follow-up response was more conservative (i.e. shared less information), while a smaller absolute variation means more consistent results.

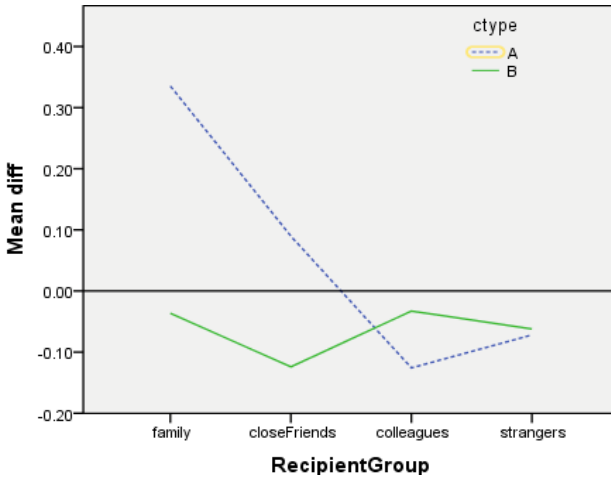


Fig. 2. Means difference in participants' responses (y-axis), grouped by recipient type (x-axis) and system type. Note that system A (dark blue) had no trajectory reminders, while system B (light green) had trajectory reminders.

5 Discussion

The results show that trajectory reminders (system B) increased the consistency of location sharing preferences elicited using a retrospective protocol, with users of System B being about twice more consistent in their responses. This suggests that trajectory reminders, and possibly other types of contextual cues, can significantly improve the reliability of location sharing retrospective protocols. It should be noted that both systems had a set of common reminders acting as contextual cues. These were, for example, the pinpointed location on a map that appeared in both systems, the names of the nearby streets, and the location of significant nearby landmarks. Hence, the difference in consistency across the two conditions is not due to the presence of reminders, but to the addition of an extra reminder in System B, namely trajectory. Therefore, it is expected that the data collection mechanism increased participants' consistency across both conditions, thus possibly masking the true effect of the trajectory reminders.

Orthogonal to the effect of trajectory reminders, the study showed a significant effect of recipient group on the consistency of participants' responses. Specifically, participants were most inconsistent in their preferences regarding family members, showing a tendency to increase the amount of information they chose to reveal to family members in the second questionnaire by about 0.13 points. On the other hand, participants' responses regarding the other recipient groups varied by up to 0.03. This suggests that participants found it more challenging to reconstruct their experience and thus accurately recall their sharing preference in relation to close family members. An explanation for this finding is that participants tend to behave in pre-determined manner when deciding how to share their location with non-family members, while not so for family members, thus leading to increased inconsistency.

It should be noted that family members mean different things to different people, and family relationships can vary significantly. To minimize this discrepancy, the study required participants to give names of specific people to act as potential recipients in the questionnaire. This ensured that across the two questionnaires individual participants were asked questions about precisely the same potential recipients rather than, for example, "a family member". Clearly, however, the differences between participants' perception of family relationships may vary. Additionally, some participants noted that their family members live in a different city and hence had no reason to share their exact address with their family while they were in the city where this study was conducted. Their exact location within this city did not mean much to their family and hence did not practically reveal any useful information in addition to the city-level granularity.

6 Implications

An important implication for designing systems supported by the results is that the incorporation of trajectory reminders increases the consistency in participants' stated preferences. This implication is directly applicable to systems that employ mechanisms such as auditing and learning from the user [7]. Such systems, for example, allow users to examine location disclosure events that they or the system made, and indicate whether they are acceptable or not.

In addition to the implications for designing systems, the results presented here have important implications for designing studies. The methodological difficulty in eliciting location sharing preferences is that while techniques such as the experience sampling are too intrusive, retrospective protocols suffer from the fact that they may introduce some unreliability in the elicited preference data. The results of this study show that one mechanism by which such unreliability can be substantially reduced is by incorporating trajectory reminders when eliciting preference data.

The scope of this study, and its assessment of trajectory reminders, is strongly focused on the domain of location sharing applications. However, trajectory reminders may be themselves useful in studies unrelated to location sharing, but rather considering context-aware systems. Since location is an integral element of context, trajectory reminders may be used to help users reconstruct the context for a specific event for which they need to express a preference. Examples include studies that require users to recall preferences regarding, for example, a smartphones' behavior. It would be interesting to explore if the effect of trajectory reminders observed in this study would hold in a such a different context.

7 Limitations

It is important to keep in mind the context in which the results of the effect of trajectory reminders presented in this study were observed. The sample of participants comprised mainly of young students and staff (median age = 27), mostly males, from a university, and care must be taken while trying to interpret the implications of these results to broader demographics. In addition, our study did not account for the

differences in the recipient groups such as “family” between participants in the two conditions. Indeed, in an informal follow up interview one participant mentioned that his family members lived in a different city and hence it would have made no difference to his family whether he shared his exact location or just his city level location with them during the period of the study. Future work in understanding the effects of trajectory reminders must address these issues.

8 Conclusion

This paper demonstrates that trajectory reminders, which are a type of contextual cue, can help elicit more reliable responses from participants in retrospective protocols that collect location sharing preferences. The study shows that indicating the locations visited before and after a specific visited location can elicit more reliable location sharing preferences for that location visit. It is argued that trajectory reminders help participants reconstruct more accurately their experience of the location visit in question, and therefore provide more reliable stated preference responses.

Acknowledgements. This work is funded by the Portuguese Foundation for Science and Technology (FCT) grant CMU-PT/SE/0028/2008 (Web Security and Privacy).

References

1. Barkhuus, L.: Privacy in location-based services, concern vs. coolness. In: *Mobile HCI 2004 Workshop: Location System Privacy and Control* (2004)
2. Benisch, M., Kelley, P.G., Sadeh, N., Cranor, L.F.: Capturing location-privacy preferences: Quantifying accuracy and user-burden tradeoffs. *Personal and Ubiquitous Computing*, 1–16 (2010)
3. Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 106(36), 15724–15728 (2009)
4. Hektner, J.M., Schmidt, J.A., Csikszentmihalyi, M.: *Experience Sampling Method: Measuring the Quality of Everyday Life*. Sage Publications, Thousand Oaks (2007)
5. Hong, J.I., Landay, J.A.: An architecture for privacy-sensitive ubiquitous computing. In: *MobiSys 2004*, pp. 177–189 (2004)
6. Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., Stone, A.A.: A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science* 306(5702), 1776–1780 (2004)
7. Kelley, P.G., Hanks Drielsma, P., Sadeh, N., Cranor, L.F.: User-controllable learning of security and privacy policies. In: *AISec 2008*, pp. 11–18 (2008)
8. Khalil, A., Connelly, K.: Context-aware telephony: privacy preferences and sharing patterns. In: *CSCW 2006*, pp. 469–478 (2006)
9. Robinson, M.D., Clore, G.L.: Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin* 128(6), 934–960 (2002)
10. Schwarz, N., Kahneman, D., Xu, J., Belli, R., Stafford, F., Alwin, D., et al.: Global and episodic reports of hedonic experience. *Calendar and Time Diary Methods in Life Course Research: Methods in Life Course Research*, p. 157. Sage Pubns., Thousand Oaks (2008)

11. Tsai, J.Y., Kelley, P., Drielsma, P., Cranor, L.F., Hong, J., Sadeh, N.: Who's viewed you?: the impact of feedback in a mobile location sharing application. In: CHI 2009, pp. 2003–2012 (2009)
12. Wagner, D., Lopez, M., Doria, A., Pavlyshak, I., Kostakos, V., Oakley, I., Spiliotopoulos, T.: Hide And Seek: Location Sharing Practices With Social Media. In: MobileHCI 2010, pp. 55–58.
13. Tulving, E.: Episodic Memory: From Mind to Brain. *Annual Review of Psychology* 53(1), 1–25 (2002)
14. Karapanos, E., Martens, J.-B., Hassenzahl, M.: Reconstructing Experiences through Sketching. Arxiv preprint, arXiv:0912.5343 (2009)
15. Hsieh, G., Li, I., Dey, A., Forlizzi, J., Hudson, S.E.: Using visualizations to increase compliance in experience sampling. In: Proceedings of the 10th International Conference on Ubiquitous Computing (2008)

Navigation Time Variability: Measuring Menu Navigation Errors

Krystian Samp and Stefan Decker

Digital Enterprise Research Institute, National University of Ireland, Galway
{krystian.samp, stefan.decker}@deri.org

Abstract. The subject of errors in menu studies is typically limited to reporting error rates (i.e., the number of clicks missing target items) or even completely neglected. This paper investigates menu navigation errors in more depth. We propose the *Navigation Time Variability* (NTV) measure to capture the total severity of navigation errors. The severity is understood as time needed to recover from the errors committed. We present a menu study demonstrating use and value of the new measure.

Keywords: Navigation Time Variability, errors, navigation, menus.

1 Introduction

There has been a great deal of research into alternative menu designs, including numerous cascading menu improvements [1, 4, 7, 12], radial menus [11], marking menus [8], and much more.

Research studies of menu designs typically focus on measuring performance in the form of selection times and error rates in the form of the number of clicks missing the target menu item. Sometimes the subjective perception is also measured with a post-experiment questionnaire.

When it comes to errors, it turns out that the error rates are low, typically below 5% (e.g., [1, 4, 5, 11, 12]). The error rates are often not significantly different for tested designs. Some studies abandon the analysis of errors, concluding that the errors are too sparse to provide any interesting insights (e.g., [5]). These results would suggest that errors do not play an important role in menu selection.

It is important to note, however, that the traditional way of measuring errors focuses only on one particular type of navigation error—i.e., clicks missing a target item. However, navigation does not consist only of mouse clicks but of all motor actions that the user needs to perform when selecting from a menu (e.g., dwelling or moving the mouse pointer). These other actions can result in errors which are not captured by counting the number of incorrect mouse clicks. For example, in the cascading menu small steering errors causing incorrect selection changes or unexpected sub-menu disappearance do not increase number of incorrect mouse clicks. Navigation is an important part of menu selection, concerning both novices and experts [3]. Therefore, it is important to understand the problem of menu navigation errors more thoroughly.

In this paper we propose a simple measure of navigation errors in menus. The data required by the measure is typically collected in menu studies. Our goal is to describe the measure and demonstrate its use and value in a traditional menu study.

The rest of the paper is structured as follows. First, we present the related research. Next, we describe the measure. The menu study follows where we demonstrate the use and value of the measure. The paper finishes with the conclusions.

2 Related Research

There are three components of menu selection: visual search, decision, and navigation [3]. Novices search for an item and then navigate to it. Experts, on the other hand, do not have to search for menu items as they know their locations. They decide which item to select and then navigate to it. If the menu is hierarchical, search and navigation (novices) or decision and navigation (experts) are performed multiple times for each menu level.

Navigation concerns novices and experts. It refers to all motor actions that the user needs to perform when selecting from a menu (e.g., moving the mouse pointer, dwelling, mouse clicks). Navigation is an important part of menu selection. Novices and experts can spend between 20-50% of the total selection time on navigation [11].

There are many possible differences in how menus are navigated. Some menus employ a point-and-click interaction style [11], others a dragging interaction style [8]. Some menus restrict navigation trajectory when moving between the sub-menus, others do not [4, 12]. Some menus make the navigation to frequently selected items easier by dynamically increasing their sizes [3]. Some menus facilitate navigation between the sub-menus; for example, by attracting the cursor towards an open sub-menu [1] or by opening a sub-menu faster if the cursor moves towards it [7]. These are only few examples. The plethora of menu navigation techniques makes it important to understand how error prone they are, beyond the number of clicks missing target items.

The importance of navigation errors is informally established in the HCI community. Numerous cascading menu improvements are motivated by various navigation problems [1, 4, 7, 12]. Pastel [10] shows that steering through sharp corners, like in the case of the cascading menu, can induce errors. Kobayashi [7] provides empirical evidence that the number of unexpected submenu appearances, again for the cascading menu, can be substantial. These works hint at some navigation problems beyond traditional error rate but are quite specific in what they focus on (i.e., one type of error particular for the cascading menu). None of the work has focused on a general approach to measuring the total impact of all navigation errors.

To address this problem, an approach focusing on measuring errors of pointing devices could be considered. MacKenzie et al. [9] proposed seven accuracy measures to elicit differences among pointing devices in precision pointing tasks :1) target re-entry 2) axis crossing, 3) movement direction change, 4) orthogonal direction change, 5) movement variability, 6) movement error, and 7) movement offset. The measures, however, do not necessarily represent menu navigation errors. Axis-crossing or movement changes do not have to result in errors such as sub-menu disappearance or selection changes – i.e., errors that the user needs to correct. This will depend on

particular menu design and size of the committed error. The above measures are well suited to assess deviations from an optimal pointing solution. It is arguable, however, if such a general optimal pointing solution exists for hierarchical menus (e.g., should the steering finish at the border of a sub-menu, in its center, or somewhere else?).

The above approach and other approaches based on counting specific types of errors also pose demands on the experimental software. All the possible errors have to be tracked individually. This might be difficult in some environments such as those employing third-party applications or toolkits.

We aim at creating a simpler and more abstract approach. We want to refrain from listing a priori all possible types of navigation errors and tracking these in the experimental software. Our goal is to assess the total impact of all navigation errors, not contribution of the individual predetermined types of errors. Such an approach would enable more immediate view of the problem of navigation errors in menus.

3 Navigation Time Variability Measure

There are many possible sources of navigation errors stemming from the many different ways of navigating menus (some examples mentioned in section 2).

We propose the following view of the problem of navigation errors: A user trying to be as fast and as accurate as possible should be able to navigate to the same target item multiple times within a similar time frame. An increased variability in navigation time for the same target item indicates navigation errors. This is because recovering from navigation errors (e.g., re-pointing to a target item) requires additional time which is not strictly related to navigating towards a desired item.

Note that the increased variability is not necessarily connected with an event such as sub-menu disappearance. The corrections done by the users to prevent the errors—for example, a temporary change of speed-accuracy strategy preventing sub-menu disappearance—will also increase variability. This is a desired behavior because such corrections also indicate navigation difficulties which are the object of our interest.

According to the above view, what we focus on is not the occurrence of a particular event indicating an error but rather the occurrence of variability in navigation time indicating extra time spent on recovering from errors. Consequently, it is not we who decide where the navigation error occurred, but rather the user by making necessary, time-consuming corrections.

The above view of navigation errors takes into account severity of the errors. Severity is an important aspect of errors having a strong effect on user perception [6]. In our case, the more severe navigation errors require more time to recover and thus lead to larger variability. This is important because we can expect different types of navigation errors to cause different degree of difficulties.

To formalize the described variability we propose the *Navigation Time Variability* measure (NTV). It is calculated as follows. (1) For each participant we establish the min-max range of navigation times obtained for **the same menu item** with the same menu design. We did not use standard deviation instead of min-max as we expect small number of measurements per item per participant (i.e., two or three—we discuss this further in section 3.1). Min-max range also assures easy interpretation of the results as it represents the difference between the fastest and the slowest navigation to

the same menu item. (2) For each participant, we calculate the average of the min-max ranges across the different menu items. The result is one value per participant representing a single NTV score.

Since the NTV is computed on per participant basis, the inferential statistics can be used to seek significant differences between tested menus.

The NTV measure has four important characteristics. First, the sources of navigation errors do not have to be known in advance. Second, it focuses on all the errors which truly require user time and effort to be corrected. Third, the measure is based on severity of the errors—i.e., more severe navigation errors result in higher scores. Fourth, the measure assures meaningful interpretation—i.e., it expresses, in a statistical sense, how much additional time is required to correct the committed errors. In light of these characteristics, it becomes clear that the measure is not suited to identify sources of the errors or the contribution of the predetermined types of errors. Rather, its goal is to assess and compare the total impact of the navigation errors.

3.1 Measuring Navigation Times – Practical Considerations

The navigation times have to be collected in an experiment. However, the experimental task cannot consist only of navigation. Consequently, some variability in collected times (i.e., menu selection times) might be also attributed to other components present in the experimental task. For example, if the task requires a simple decision apart from navigation, the variability of the decision time will contribute to the total measured variability. Therefore, if the goal is to assess navigation errors, it is important to use a task that emphasizes navigation and minimizes contribution of the other components (i.e., the other components should be relatively small compared to navigation and have small variability ranges). In particular, the menu selection task should not require visual search or problem solving because both can take long and introduce extensive variability.

If the task emphasizes navigation, then the NTV computed on selection times will allow one to assess qualitatively which menu causes fewer navigation errors. However, if the goal is to assess the exact quantitative extent of the navigation errors, the contribution of the other components have to be factored out. To this end, the components have to be known and have known variability ranges.

A considerable amount of menu studies have focused on measuring navigation performance. They employ a common task which emphasizes navigation and adheres to the above characteristics. The task is to select an item from a single or hierarchical menu. All items in a selection path (i.e., parent items and the target item) are highlighted. Consequently, the participants do not have to search for each item nor decide which item to select. The participants only need to: 1) respond to visual stimuli and prepare the movement which takes $234 \pm 41\text{ms}$ [2] and then 2) navigate to the item. Ahlstrom [1] demonstrated empirically that the above task emphasizes navigation. He accurately modeled total selection times using only navigation component based on the Fitts' and the steering pointing laws.

The above task is commonly used in menu studies (e.g., [1, 4, 12]). It is also common to administer two or more blocks of menu selections for each tested menu design

(e.g., [1, 3]). Therefore, there is potential for similar menu studies to use the proposed measure without modifying the experimental design but merely by extending the error analysis part.

4 Experiment

The objective of this section is to demonstrate use and value of the NTV measure in a traditional menu study employing the navigation task presented in section 3.1.

4.1 Menus

Two menus were used in the experiment. The first menu was the Cascading Menu (CS), which we chose because it is known to cause navigation errors [1, 4, 7, 12]. The implemented CS menu was equivalent to those found in contemporary applications such as MS Word. In short, a mouse click was used to open the menu and select the final item and dwelling over a parent item for 1/3 seconds opened a sub-menu. Moving between the sub-menus required steering, the sub-menus did not disappear immediately upon steering errors but after short delay, and the item under the cursor was highlighted blue.

The second menu was the Compact Radial Layout menu (CRL) [11] which we chose because it was designed specifically to decrease navigation errors by the means of eliminating dwelling, eliminating steering, decreasing navigation distances, increasing item sizes, and not restricting navigation paths. The implemented CRL menu was equivalent to that presented in [11]. In short, a mouse click was used to open a menu, select a parent item, and the final item. The items were represented by circles, the levels by concentric rings. The innermost ring corresponded to the first level and again the item under the cursor was highlighted blue. Figure 1 shows examples of selection sequences for both menus.

The size of the CS menu items was set to 215×19 pixels because this size could accommodate three word labels. The diameters for the CRL menu items were 44, 48, and 52 pixels for the first, second and third level respectively. The resulting circles provided enough space to also accommodate three word labels (eventually broken to two or three lines) while avoiding overlaps.

The menu content consisted of 6 items on the first level, 11 items on the second level, and 15 items on the third level. We used one to three character labels to help the user visually separate the menu items; however, they did not have any meaning.

4.2 Measures

The study included quantitative and qualitative measures. Our main dependent variables were selection time and the number of mouse clicks missing target items. In the post-experiment questionnaire, the participants were asked to rank each of the menus on a 1-7 Likert scale according to the following criteria: *Error prone*, *Frustration*, and *Ease of use*.

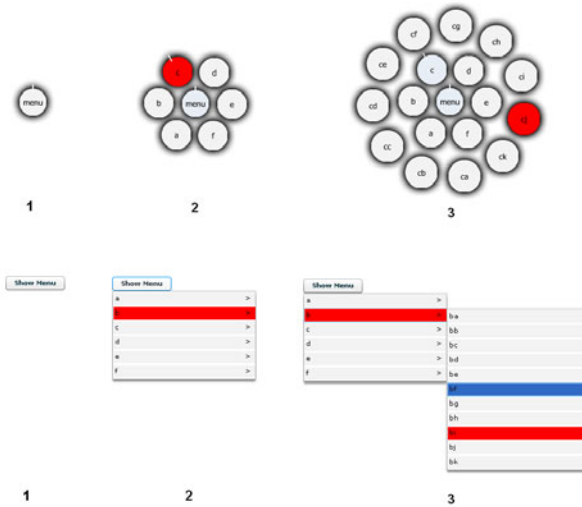


Fig. 1. Selection sequence for a target item on the second level for the CRL menu (at the top of the figure) and the CS menu (at the bottom). All items in a selection path are highlighted red.

4.3 Procedure

A total of 28 participants took part in the experiment (8 female and 20 male). All were between ages of 20 and 28. The navigation task presented in section 3.1 was used.

The procedure was as follows. (1) Participants were told the procedure of the experiment. (2) For both menus the system provided a two sentence description of their behavior and allowed the participants to practice for two minutes. (3) The experiment started. The participants were asked to complete the selections as fast and as accurately as possible. For each menu, each participant completed two identical blocks of 45 menu selections (separated by a two minute break): 10 selections on the first level, 15 on the second level and 20 on the third (with a one minute break between the levels). The item sequences were generated randomly. The menu ordering was balanced. The time was measured between the first click opening the collapsed menu and the final click on the target item which collapsed the menu. (4) The participants completed the post-experiment questionnaire.

4.4 Results and Discussion

The traditional measure of error rate indicated that 1.6% of the trials were erroneous (i.e., a mouse click missing a target item) for the CRL menu and 3.7% for the CS menu. Because the scores of error rates were sparse and not normally distributed we used the Wilcoxon Signed-Rank test for statistical analysis. The test indicated that the error rate is not significantly different for both menus ($p > 0.05$).

The results of the questionnaire were as follows. All the participants claimed that they were more error prone with the CS menu. 18 participants were more frustrated with the CS menu, 4 with the CRL menu, and 6 did not see any difference. 6 participants considered the CRL menu easier to use, 5 the CS menu, and 17

considered both menus equal. Because the subjective scores are non-parametric, we again used the Wilcoxon Signed-Rank test. The CRL menu is perceived as significantly less erroneous ($p < 0.01$) and less frustrating ($p < 0.01$) than the CS menu. No difference was found with respect to ease of use.

The results indicate that the CS menu is strongly perceived as more erroneous and more frustrating. However, this perception cannot be attributed solely to the error rate (i.e., the number of clicks missing target items) as it was not significantly different for both menus. The subjective perception is not supported quantitatively. The results hint that there is more to navigation errors than clicks missing their target items.

Using data from two blocks, we calculated the NTV for each participant. Table 1 shows the summary of the results.

Table 1. The NTV on the three menu levels averaged across the participants. The NTV scores followed normal distribution within each menu x level cell.

	Level 1	Level 2	Level 3	Marginal mean
CRL MENU	85 ms	164 ms	232 ms	160 ms
CS MENU	176 ms	862 ms	1005 ms	681 ms

To determine if the NTV differences are significant, we performed pair-wise comparisons using dependent measures t-test. The family-wise significance level was adjusted according to the number of tests performed. The comparisons revealed that the menus were not different on the first level but the CRL menu generated lower NTV than the CS menu on the second and the third level. For the CS menu, each level produced higher NTV compared to the previous levels. For the CRL menu, the first level had lower NTV than the second and third levels which did not differ between themselves. All the reported differences are at level $p < 0.001$.

The CRL menu. For the CRL menu, the NTV on the first level is the lowest (85ms). It can be attributed to the variability of reaction time of aimed movements (82ms [2]). The NTV on the second and the third level are doubled and tripled respectively as these levels require one additional reaction time compared to the previous level. There is no sign of any additional navigation variability which could indicate navigation errors. The participants were able to maintain roughly constant navigation performance on each level when selecting the same items.

The CS menu. For the CS menu, the NTV rapidly decreases between the levels. This indicates that participants, who performed well with the CRL menu, had more problems with the CS menu. The size of the NTV for the CS menu and the size of the differences in the NTV between both menus certainly cannot be attributed to the variability of reaction time of aimed movements. Furthermore, as the NTV for both menus is similar for the first level, we conclude that it is the navigation between the levels that causes rapid increase of the NTV for the CS menu. After factoring out the variability of reaction time of aimed movements¹ from the NTV scores in Table 1, we estimate that recovering from the navigation errors takes approximately 94ms, 698ms, and 759ms for selections on the first, the second and the third level respectively. These amounts are substantial compared to the typical total selection times—e.g.,

¹ Reaction time being 82ms [2] for the first level selections, 2×82 ms for the second level selections, and 3×82 ms for the third level selections.

according to [12], approximately 1750ms and 2500ms for the second and the third level selections. This hints at the importance of navigation errors.

The above results provide quantitative support for the subjective findings regarding errors. To remind the reader, the CS menu was perceived as significantly more erroneous (stated unanimously) and significantly more frustrating than the CRL menu. In contrast to the traditional error rate, the NTV captured this difference.

Finally, the results also demonstrate that the NTV can be small, even for selections on the third level. This finding is important because it supports the underlying assumption of the NTV measure stating that a user can navigate to the same menu item within a similar time frame.

5 Conclusions

Our goal in this paper was to describe a new measure of navigation errors in menus and show its use in a traditional menu study. We demonstrated that the proposed NTV measure gives quantitative information on navigation errors beyond traditional measure of error rate. It also supported the qualitative findings.

The measure is not intended to replace the traditional measures (e.g., error rates). Rather, we consider it supplementary measure, with the potential to assess the total impact of navigation errors. The new measure increases the theoretical knowledge base on differences between menu designs.

References

1. Ahlström, D.: Modeling and improving selection in cascading pull-down menus using Fitts' law, the steering law and force fields. In: CHI, pp. 61–70 (2005)
2. Bekkering, H., Adam, J.J., Kingma, H., Huson, A., Whiting, H.T.A.: Reaction time latencies of eye and hand movements in single- and dual-task conditions. *Exp. Brain-Res.* 97 (1994)
3. Cockburn, A., Gutwin, C., Greenberg, S.: A Predictive Model of Menu Performance. In: International Conference on Human Factors in Computing Systems (CHI), pp. 627–636 (2007)
4. Cockburn, A., Gin, A.: Faster cascading menu selections with enlarged activation areas. In: Graphics Interface Conference (GI), pp. 65–71 (2006)
5. Ellis, J., Tran, C., Ryoo, J., Shneiderman, B.: Buttons vs. menus: An exploratory study of pull-down menu selection as compared to button bars. Technical Report, vol. 10 (1995)
6. Feng, J., Sears, A.: Beyond errors: measuring reliability for error-prone interaction devices. *Journal Behaviour & Information Technology* 29(2), 149–163 (2010)
7. Kobayashi, M., Igarashi, T.: Considering the direction of cursor movement for efficient traversal of cascading menus. In: UIST, pp. 91–94 (2003)
8. Kurtenbach, G.: The Design and Evaluation of Marking Menus. PhD Thesis (1993)
9. MacKenzie, I.S., Kauppinen, T., Silfverberg, M.: Accuracy measures for evaluating computer pointing devices. In: CHI, pp. 9–16 (2001)
10. Pastel, R.: Measuring the difficulty of steering through corners. In: CHI, pp. 1087–1096 (2006)
11. Samp, K., Decker, S.: Supporting menu design with radial layouts. In: International Conference on Advanced Visual Interfaces (AVI), pp. 155–162 (2010)
12. Tanvir, E., Cullen, J., Irani, P., Cockburn, A.: AAMU: adaptive activation area menus for improving selection in cascading pull-down menus. In: CHI, 1381–1384 (2008)

Challenges in Designing Inter-usable Systems

Ville Antila¹ and Alfred Lui²

¹ VTT Technical Research Centre of Finland, Kaitoväylä 1, Oulu, Finland
ville.antila@vtt.fi

² Fjord, 19 Margaret Street, London, United Kingdom
alfred.lui@fjord.co.uk

Abstract. Interactive systems are increasingly interconnected across different devices and platforms. The challenge for interaction designers is to meet the requirements of consistency and continuity across these platforms to ensure the inter-usability of the system. In this paper we investigate the current challenges the designers are facing in the emerging fields of interactive systems. Through semi-structured interviews of 17 professionals working on interaction design in different domains we probed into the current methodologies and the practical challenges in their daily tasks. The identified challenges include but are not limited to: the inefficiency of using low-fi prototypes in a lab environment to test inter-usability and the challenges of “seeing the big picture” when designing a part of an interconnected system.

Keywords: Interaction design, cross-platform systems, inter-usability.

1 Introduction

While the amount and diversification of computing devices is increasing the boundaries between the systems that can be accessed with them is blurring. Users expect to have access to the same applications and services with a multitude of devices and expect that this process is highly optimized to support the capabilities of the device at hand. In addition to this optimization, applications are expected to support efficient flow of interaction and coherence across the different user interfaces [10]. Today’s products are becoming increasingly ubiquitous systems, hybrids of hardware, software and services [2]. It is not yet very clear what design and evaluation techniques should be used to design these products.

Nevertheless, these are the challenges that designers of applications and services are facing today. In this paper we take a look at the challenges in designing *inter-usable* systems from the designers and developers perspective. With inter-usability we mean the usability and user experience across the different user interfaces of a given system (also including inter-device interactions, similarly to [1]). The added complexity in the system requirements translate to added complexity in interaction design as well as challenges in evaluating the overall usability of the system.

To identify some of the specific challenges that practitioners of interaction design are facing, we conducted semi-structured interviews with 17 specialists in interaction and user interface design, working with services and applications designed

for different levels of inter-usability. The application domains span from experimental products and prototypes mixing physical and digital elements to mass-market consumer products with user interfaces for both dedicated mobile applications and Web-based services. We asked the designers of these systems about the design and development processes they use, what kind of methodologies they use and what are the main challenges they face in their daily work. Based on the interviews we analyzed the specific challenges and situated them to the different stages of the design process to be able to refine the requirements of emerging design methodologies and tools.

2 Background

We are starting to see an increasing amount of interconnected devices and services that will require new design methodologies and new kind of understanding to make them (inter) usable [3]. The understanding of challenges in designing interconnected and inter-usable systems as opposed to traditional interaction design has recently caught more attention because of the proliferation of cloud-based, multiple touch-point service. Denis and Karsenty [1] propose a conceptual framework for inter-usability. The design principles address inter-device consistency, transparency and adaptability as well as knowledge and task continuity. They describe continuity in cross-platform systems as shared memory between the system and the user. In an inter-usable system the user can recover the state of operation and continue the task after a transition from one device to another. Similarly, Wäljas et al. [10] introduce a conceptual framework for cross-platform user experience. They argue that the central elements for cross-platform user experience are:

- *Composition*: How different platforms (applications and devices) within a system relate to each other. This can be about role allocation of different devices, distribution of functionality or modularity of functionality
- *Continuity*: Interoperability is supported by carrying out transitions between platforms, e.g. seamless synchronization of data and content, explicitly supporting users in migrating tasks across platforms by proposing linkages to other devices in the system
- *Consistency*: Keeping the user experience coherent across multiple platforms to assist the Continuity element of the cross-platform UX. This can be perceptual (look and feel), semantic (symbols and terminology) and syntactic (interaction logic).

Inter-usable systems are composed of applications designed for different platforms. The different applications can support different activities within the system and can interact with different service "touch-points". The relation between these applications can have different requirements depending on whether the actual service is just multi-channeled through these applications or whether the different functionalities are actually distributed amongst them (whether the different platforms have a distinct purpose in the system) [8].

The current methodologies for designing and evaluating interactive systems are usually limited to evaluating systems on a single platform or device, and do not usu-

ally take into consideration the inter-usability aspects of cross-platform consistency and continuity. To address the challenge of designing user interface for multiple platforms there has been ongoing research on conceptual frameworks (metaphors and interaction styles) [7], design frameworks (patterns, widgets, UI component toolkits) [5, 6], application frameworks (both design-time and run-time support) [4] and validation frameworks (usability techniques for testing multiple user interfaces) [7]. However, often there's a gap between frameworks, available methods and tools, and the practical needs of practitioners [9]. In this paper we take a look at the specific challenges and requirements for the methods and tools from a practitioner's perspective and situate the current state of the art based on these requirements.

3 Research Approach

To identify current challenges of interaction design, we carried out semi-structured interviews with 17 professionals from 10 different organizations. The organizations spanned from major consumer product manufacturers to small design firms and start-ups. The interviewees included 6 interaction designers, 4 researchers in the field of user-centered design, 2 freelancer designers, and 3 founders of start-ups in the field of ubiquitous computing as well as one application engineer and one technical director. The interviews were conducted face to face and over the phone and each interview lasted about one hour. The interviewees were recruited through personal connections and were selected based on the knowledge of their work. The criterion for selection was that the work includes components, which are interconnected with some level of measured usability.

During the conversations we first asked the interviewees about their background, the type of company they were working for, what kind of products they were designing and the typical users of those products. We also asked the interviewees about the common design models and methodologies they use in their work. We subsequently probed into the possible challenges and difficulties they face in applying these methodologies in their work (regarding the suitability of methodologies for the given tasks). In addition to the structured questions we developed follow-up conversations to interesting answers, especially regarding the challenges the interviewees were facing.

All interviews were recorded and transcripts were made from the recordings. The transcripts were coded using a set of themes: *design process*, *usage of UI methodologies* and *biggest challenges*. The challenges were also coded into clusters based on the design process phase. The individual challenges were also clustered together to find common themes of challenges. We describe the main findings and analysis in the following section.

4 Findings

The most used UI design methodologies identified by the interviews were *iterative design* and *rapid prototyping*. Nevertheless when probing to the actual practices they seem to depend heavily on the team and company size. In small design firms, which mostly work on custom prototypes and products, the emphasis is on rapidly

implementing ideas, which then translates to minimum use of traditional user-centered research practices and formal usability testing. In bigger companies targeting products for mass-markets, the use of rigorous processes usually translates to well-defined use of UI design methodologies. In the latter cases iterative and rapid prototyping is preferred, since it allows them to have early user involvement and avoid UI design changes in later phases when major changes to the product design is usually too late. We found this same division also applicable to the identified challenges (the identified challenges are listed in Table 1).

Based on the interviews we found a set of challenges that the designers were facing in their work. Some of these challenges were quite universal, such as the challenges in time and communication. Therefore, those are not discussed further. In this paper we concentrate on the following identified challenges related to inter-usability:

- *Development environment constraints*: The limitations of current development environments (e.g. challenges in prototyping systems which include both hardware and software)
- *Domain restrictions*: The restrictions of applications domains and their implications to practical choices between technologies (e.g. deployment strategies in restricted domains like healthcare)
- *Acquiring domain knowledge and user feedback*: Challenges in acquiring the right domain knowledge and user feedback early enough (e.g. challenges in simulating systems which span different devices with different capabilities)
- *Developing for multiple platforms*: The challenges for multiple platforms span from specific challenges of developing different interactions to support different capabilities of devices to maintain the consistency and continuity of interaction in an interconnected system

Development environment constraints - In some interviews it was mentioned that the design phase and development phase should be easier to integrate. In small companies this can be done with close iteration and even using the same tools to design and prototype. In larger teams where design and implementation activities are more separated, UI component libraries can help establish this integration. This is often done implicitly, without any real integration on the tool-level. The designers often cannot test the components and UI flows interactively before the real implementation phase, which imposes difficulties to spot design faults enough early in the development.

“There are component libraries and design guidelines between product families, but these do not go all the way into the ‘tools’ –level. It would be useful to share the same common tools with developers to avoid the usual challenge [the designers face] ‘this cannot be implemented on this platform’” – Interaction designer, UI designs for mobile and Web-based applications.

While these challenges are also observed in traditional design and development workflows, the lack of standard library and data protocols across platforms compound their impact on the efficiency between the teams.

Domain restrictions - Some domains impose restrictions on the used technologies by using old versions of software platforms or by limiting the capabilities of installing new software. Also in some fields expert insights and inputs are crucial for a successful design but the inputs are hard to get because of tight schedules or policies. The limitations on the deployment of new technologies affect both the continuity and consistency of cross-platform solutions in that domain. This creates implications to the used technologies. Following example describes some of the challenges imposed by the healthcare domain:

“Many of users work in an environment where installation of new software or packages on their computers is restricted, for example by hospital IT departments. We have to work within these restrictions. For that reason we choose web based UI’s for most of our projects because it does not require installation of software on the [devices] of the users” – Application engineer, user interfaces for applications in the domain of healthcare.

Acquiring domain knowledge and user feedback - Knowing the users and their needs is a major challenge identified in all interviews. Nevertheless the type and way of getting the knowledge from users may differ greatly by the type of product. In experimental projects by small companies the target group can be hard to define and involve in the early stage so the design process is very much driven by intuition and personal insights. In specific domains like health-care the challenge often is to find out what the real user needs are. This can be made difficult by the time restrictions of domain professionals.

When designing an interconnected system another challenge is the inefficiency of user testing in lab environment. In many cases the interviewees identified early user feedback as important or vital, but also noted that they cannot get feedback until they have a functional prototype which can be tested in the field. Testing usability of solutions to accommodate change in context is particularly hard due to the lack of standardized research framework and analysis methods.

“A lot of the things we are changing, we will do more of a [...] we put it out there and see if people are using it, we don’t spend that time doing a full usability type of thing. Moving towards what the ‘Google model’ is like, they almost put up the lab version or they have a beta version and people just use it, get the feedback and just improve it.” – Co-founder of a start-up on professional sports tracking technology.

Targeting multiple platforms - The challenge in designing for multiple platforms is the different capabilities and user interaction metaphors they may have. There is no easy way to meet this challenge. Part of the design always has to be adjusted to the target device on its own, but re-usability of user interfaces across the platforms is also needed for a unified look-and-feel when dealing with an interconnected system.

“As [the application] must run on several devices, and the devices have different capabilities in terms of display and user controls, it is difficult to design a UI that can be used as fast and easy on all the devices.” – Technical director, mobile and ubiquitous gaming applications.

Table 1. Identified challenges grouped by the design process phases

<i>Process phase</i>	<i>Identified activities</i>	<i>Identified challenges</i>
Early phase	Iterative concept creation and evaluation, Interviews, Contextual inquiries, Focus groups	<ul style="list-style-type: none"> • Hard to evaluate concepts without functional prototypes (simulation of interactive system) • Seeing the big picture when designing a part of a interconnected service (often designers just focus on part of the system which can create inconsistencies and discontinuities between the parts)
Design phase	Rapid prototyping (e.g. paper), Integration of design and development tools, Task analysis	<ul style="list-style-type: none"> • Each failed experiment with physical objects incurs material, labor and transportation costs (unlike with fully digital products/services)
Development phase	Functional prototyping, Use of (in-house) component libraries	<ul style="list-style-type: none"> • None of the tools available today is sufficient to build and test inter-usable systems • Basic tools such as IDEs, Flash and PCB design tools are generic enough to fill the gap but by no means efficient for designers who want to weave digital data into physical materials
Evaluation phase	Expert reviews, Usability testing, Field trials	<ul style="list-style-type: none"> • User testing of embedded devices and interconnected services using low-fi prototypes in a lab environment is inefficient • Difficult to evaluate the whole (interconnected) system; evaluation of separated parts does not necessarily correspond to good overall (inter) usability

5 Analysis and Implications for Design Methodologies

Based on the information gathered from the interviews we analyzed the possible needs for refinement of methods and tools. We identified three basic types of needs from the interviews. There is currently a need for better *integration between the design and development tools*. A number of interviewees identified that it is challenging to test designs on different platforms enough early in the design process. They identified a need for “mash-up” type development environment where the composition and continuity of an inter-usable system could be better tested. Another challenge, which could benefit from better tools, is the possibility to *link smaller design tasks into the “big picture”*. This can be especially difficult in systems incorporating several different user interfaces for different devices and platforms. This *cross-platform design* support also needs different methodologies than more traditional single platform systems. The current design methodologies do not take into the consideration the need to support effective transitions and interaction metaphors across platforms. The implications of findings are described more in the Table 2.

Table 2. Implications of findings to requirements for design methods and tools

<i>Identified need</i>	<i>Description</i>	<i>Requirements for methods and tools</i>
Support for “seeing the big picture” – how the design fits in the whole system	Consistency and continuity is important in cross-platform user experience, the current design methods and tools do not offer support for evaluating this	<ul style="list-style-type: none"> • Early prototyping through simulation • Evaluation metrics to test consistency (semantic and syntactic) and continuity in cross-platform and cross-device interactions
Integration between design and development tools	There are challenges in integration of design and implementation. The currently available tools are not sufficient to build and test inter-usable systems (such as services with interfaces for both cloud and dedicated devices)	<ul style="list-style-type: none"> • Ability to test or “mash-up” the composition of interconnected systems (e.g. distribution and composition of functionalities between the cloud and dedicated devices) • Support for rapid prototyping
Refinement of evaluation methods and metrics to test inter-usability	There are challenges in evaluating interconnected systems early enough, and to measure inter-usability. There is also a need for early evaluation of interaction metaphors which translate between the different domains and user interfaces designed for different platforms	<ul style="list-style-type: none"> • Evaluation methods and metrics to support inter-usability, taking into account both the composition of functionalities and the continuity of interaction • Design guidelines to support semantic consistency across platforms (the use of metaphors etc.) • Ability to use efficiency measures to validate inter-usability of cross-platform interactions

The implications of our findings can be used as an indicator of growing needs for measuring and testing the inter-usability of interconnected systems. It is clear that the tools and methods in design and development at their current states do not sufficiently address these needs for cross-platform design. Based on our findings the specific needs lie in the design and evaluation phases where we found needs for both rapid prototyping tools for interconnected systems as well as better evaluation metrics to test the composition, consistency and continuity aspects of cross-platform user experience [10]. Because of that, the designing and developing for such systems currently involve a great deal of intuition and trial-and-error.

Another factor that contributes to the challenge is that devices themselves no longer offer user standardized means of manipulating information. Mobile phones, the epitome of multi-platform, cross-context devices, have morphed from a combination of a 12-digit pad, a set of 5-way navigation key and a small display to a touch-sensitive screen that can offer any number of controls. Going further, interconnected objects that are believed to embed information access into our environments in the near future can come in theoretically limitless number of forms.

5 Conclusions

In this paper we have presented the analysis and implications of findings identified through interviews of professionals in interaction design. We identified a set of specific challenges that the practitioners face on different stages of the design process. We also identified needs for further refinements of methodologies to overcome the challenges imposed by requirements of inter-usability. We acknowledge these results as preliminary and foresee a need for further research on translating the identified challenges and needs to refinements on methods and tools to support these challenges in practice. Nevertheless we see this work as an important step towards understanding the needs of inter-usability better from the point-of-view of professionals dealing with the practical challenges of designing interconnected systems and services.

Acknowledgements. This work has been supported by the SMARCOS Artemis project. We would like to thank Privender Saini, Marloes van der Hout, Josu Cobelo, Jorge Rodriguez and Guido Galeazzi for their contributions to the interviews and analysis. We also acknowledge the comments and suggestions from Tine Lavrysen regarding this paper.

References

1. Denis, C., Karsenty, L.: Inter-Usability of Multi-Device Systems – A Conceptual Framework. In: Seffah, A., Javahery, H. (eds.) *Multiple User Interfaces: Cross-Platform Applications and Context-Aware Interfaces*. John Wiley & Sons, Ltd., Chichester (2004)
2. Kuniavsky, M.: *Observing the user experience: a practitioner's guide to user research*. Morgan Kaufmann, San Francisco (2003)
3. Le Voi, H., Light, A., Rowland, C.: Towards interusability; HCI for cloud computing and embedded devices. In: *Proceedings of Designing Interaction for the Cloud Workshop in CHI 2011* (2011)
4. Melchior, J., Grolaux, D., Vanderdonckt, J., Van Roy, P.: A toolkit for peer-to-peer distributed user interfaces: concepts, implementation, and applications. In: *Proceedings of the 1st ACM SIGCHI Symposium on Engineering Interactive Computing Systems, EICS 2009*, p. 69. ACM, New York (2009)
5. Nilsson, E.: Combining compound conceptual user interface components with modelling patterns—a promising direction for model-based cross-platform user interface development. In: Forbrig, P., Limbourg, Q., Urban, B., Vanderdonckt, J. (eds.) *DSV-IS 2002*. LNCS, vol. 2545, pp. 104–117. Springer, Heidelberg (2002)
6. Seffah, A., Forbrig, P., Javahery, H.: Multi-devices “Multiple” user interfaces: development models and research opportunities. *Journal of Systems and Software* 73(2), 287–300 (2004)
7. Seffah, A., Javahery, H.: *Multiple User Interfaces: Cross-Platform Applications and Context-Aware Interfaces*. John Wiley & Sons, Ltd., Chichester (2004)
8. Segerståhl, K., Oinas-Kukkonen, H.: Distributed user experience in persuasive technology environments. In: de Kort, Y.A.W., IJsselstein, W.A., Midden, C., Eggen, B., Fogg, B.J. (eds.) *PERSUASIVE 2007*. LNCS, vol. 4744, pp. 80–91. Springer, Heidelberg (2007)
9. Väänänen-Vainio-Mattila, K., Roto, V., Hassenzahl, M.: Towards Practical User Experience Evaluation Methods. In: *International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM)*, p. 19 (2008)
10. Wäljas, M., Segerståhl, K., Väänänen-Vainio-Mattila, K., Oinas-Kukkonen, H.: Cross-platform service user experience: a field study and an initial framework. In: *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, MobileHCI 2010*, p. 219. ACM, New York (2010)

Directed Cultural Probes: Detecting Barriers in the Usage of Public Transportation

Susanne Schmehl¹, Stephanie Deutsch¹, Johann Schrammel¹, Lucas Paletta³,
and Manfred Tscheligi^{1,2}

¹ CURE –Center for Usability Research and Engineering, Modocenterstr.17/2
1110 Vienna, Austria

{schmehl, deutsch, schrammel, tscheligi}@cure.at

² ICT&S Center, University of Salzburg, Sigmund-Haffner-Gasse 18
5020 Salzburg, Austria

manfred.tscheligi@sbg.ac.at

³ JOANNEUM RESEARCH ForschungsgesmbH, Wastiangasse 6,
8010 Graz, Austria

lucas.paletta@joanneum.at

Abstract. In this paper we describe the application of a variation of cultural probing for identifying barriers in the use of public transportation for target groups with visual, cognitive or language-related handicaps. To be able to better focus on the targeted aspect - the barriers - we applied modifications to the traditional cultural probing approach: Users were encouraged to generate data related to the targeted aspect. We found that this approach can produce focused results that can be analysed fast and can help to overcome obstacles related to limitations in verbal skills or expressiveness of the user.

Keywords: User requirements, Cultural Probes, Directed Cultural Probes, elderly, immigrants, functional illiterates, public transportation.

1 Introduction

The good understanding of the users' view of the world and their needs is an essential prerequisite for designing usable products and tools. To improve quality and gathering of such information, different methodological approaches are continuously elaborated. Within HCI a qualitative method called cultural probing, has been introduced, which is used to gain insights into users living context and behaviour aspects in given contexts (such as the use of certain technology). Cultural probing benefits from enhanced user involvement, as participants are engaged to autonomously capture personal impressions from their daily life. Originally introduced by Gaver and colleagues [1], cultural probes encourage participants to generate materials in everyday situations by use of different means (e.g. camera, diary). Hence, the method allows participants to note subjective impressions using textual and graphical illustrations apart from restricted questionnaire formats. The main benefit of the concept is the richness of the user-generated data based on subjective perceptions. Whereas this approach serves to collect large amounts of

diverse information from a specific context, results from this highly subjective data need to be interpreted qualitatively. During analysis interpreters usually seek to deduce all-encompassing evidence from heterogeneous data. Similarly, they are required to process their data in ways that prevent from informational loss.

In this paper we describe the realization of a study using modified cultural probing for the identification of subjective barriers that people with special needs encounter in public transportation facilities. We report advantages and shortcomings of the new approach as well as we discuss the application possibilities and whether the method is tailored appropriately to the abilities of three user groups.

2 Related Work

Since Gaver and colleagues started doing research with cultural probes, many variations have been elaborated. Most of these variations contain diaries and photography in some form. Diaries are used to capture activities in context, to understand needs and motivations related to the use of a certain technology and to gather user requirements for design [2]. The application of the method gradually developed from an inspirational to an informal function. By conducting Technology Probes [3] existing technologies can be situated in users' homes to inspire design by exposing users to new experience. Empathy Probes [e.g. 4] have been designed to see and understand people's emotions and feelings in their natural environment, in order to support design processes. Mattelmäki and colleagues included a small diary booklet, a sheet of stickers, a disposable camera with a list of photography assignments, and ten illustrated cards with open questions. Hulkko and colleagues [5] developed a new digital contextual and self-documenting tool for studying people's actions in mobile contexts. They used mobile phones with GPRS connections, an external digital camera and a developed system for sharing and sorting the data. Crabtree and colleagues [6] proposed their Informal Probes employing various biographical approaches to encourage participants to reflect upon and articulate important personal, social, and technological features of their everyday lives. These reflections enabled designers as well as participants to formulate and elaborate the role of design in the studied culture. Until today, many variations of Gaver's cultural probes have been deployed.

Nevertheless, the richness of the collected data might also be a disadvantage in concerns of analysing the data. The traditional approach of cultural probing typically produces a huge amount of unstructured qualitative data. Additionally, it is difficult and time-consuming to identify the relevant material within this plenitude of created data. In order to differentiate from the original purpose of Gaver's concept [1] - to inspire design - and from the solely identification of user needs, we developed a variation, which asks the user to actively focus on specific problems and to detect and document relevant situations as well as to provide suggestions of how to solve the problems. We used this approach to directly detect barriers in public transportation for people with special needs.

3 Directed Cultural Probes

For the study a modified, directed cultural probing (DCP) method was elaborated. Instead of documenting the whole contextual experience, the documentation was explicitly directed to a restricted issue of the context, focused on barriers in the use of public transportation facilities. The participants – a group of elderly, a group of functional illiterates, and a group of immigrants - had to follow a standardised documentation scheme with the goal to produce focused data. The challenge was to gather and compare qualitative data from different user groups using different materials. Since participants differed in reading and language skills, the DCPs had to be adapted to their special abilities thoroughly. Hence, we will show how this approach is also suitable for contextual inquiries with other user groups.

3.1 Development

When compiling the probing packages, special requirements in relation to participants' ability to understand textual information were taken into account. In a first step a workshop with representatives of the end user organisations of three social groups was conducted on how to design the cultural probing packages in respect to specific end user needs. A generic manual served as a basis to discuss on suitability of the probes for the groups. For sending the elderly persons in the field, it was crucial that they did not have to write down things immediately. According to their physical condition it was argued that note taking in a sitting position would be more comfortable. A *diary* (instead of a *voice recorder*) was recommended in paper-pencil form to write down their experiences for elderly people and immigrants. The immigrants were allowed to complete the diary in their mother tongue, otherwise the additional effort for the participants might have kept them from documenting properly and some essential information might have gone lost.

Table 1. Different composition of the probing packages per end user group

	Elderly	Illiterates	Immigrants
Probing package	diary at home, camera, checklist	voice recorder, camera, checklist	diary, camera, checklist

Finally, workshop participants agreed that participants of all groups could document their experienced barriers by using a *disposable camera* (max. 27 pictures). The participants were asked to take a picture of the detected barrier and to comment as soon as possible on the required entries. However, according to the participants all diary entries (written or spoken) were made on the same day as the picture was taken.

Instead of the diary, the functional illiterates received a *digital voice recorder* (Olympus VN-5500) to comment the detected barriers right after taking the picture. They had a checklist on their voice recorder. To produce structured data each photograph had to be commented by a strict documentation scheme. The scheme also served as the *checklist* provided to all participants: date and time, name and brief description of the barrier, name of leg of the journey, emotional state in relation to the

barrier (5-smiley scale), behavioural reactions of the surrounding, ideas or wishes to eliminate or minimize the barrier.

The checklist was adapted for user groups: a translated version for immigrants and a version with extra-large letters for the elderly and the functional illiterates. Illustrations with pictures depicting the legs of journey should serve a better understanding by all user groups.

3.2 Participants

Three user groups were involved in the investigation: elderly people with visual and cognitive impairments, immigrants with initial lack of expertise about local language and functional illiterates with principal problems in the understanding of text. The *elderly* group consisted of 10 persons (4 men, 6 women). Their impairments varied in terms of degree of vision (e.g. nearly blind) and physical mobility (e.g. walking aid). During recruiting special prerequisites were defined: their ability to use public transportation independently, to reflect their loss of memory (in relation to dementia), and to be in a stable health condition, even if being under medication. 12 *functional illiterates* (7 men, 5 women) participated. Their mother tongue had to be German and their linguistic competency levels had to vary. The degree of variation was determined by the end user organisation. 11 *immigrants* (1 male, 10 women¹) took part in the study. At recruiting they had to establish the following preconditions: maximum duration of stay in Austria of 6 years and relatively bad German language skills. A high diversity of languages within the group of immigrants was striven for (8 different languages), to cover as much language diversity as possible.

Table 2. Total number of participants and their age (at the beginning of data acquisition)

	Elderly	Illiterates	Immigrants
number of participants	10	12	11
average age of participants (mean \pm stdv)	66,6 \pm 6,9	36,8 \pm 14,1	29,0 \pm 4,4

All participants used public transport previously. For the duration of the study, they were requested to use public transportation as often as possible, to ensure that enough data was collected. However, in order to avoid the generation of pseudo-barriers no minimum or maximum of barriers to identify was requested from participants.

3.3 Study Procedure

For the study the DCP method was combined with pre-interviews and post-focus groups. The way of how barriers were documented varied between the groups according to their special needs. After acquisition, data was reprocessed and focus

¹ Most women neither had a car nor a driving license, that's why women rather than men use public transportation. Consequently, this skewed sex ratio within the immigrant group reflects realistic demographic conditions.

groups were conducted to discuss and prioritise the detected barriers by the end user themselves.

Pre-Interview. Participants were pre-interviewed in order to get first insights: in their demography and daily habits (e.g. use of public transportation), as well as in their social environment (e.g. family situation). Furthermore, participants were asked to report on perceptions and problematic situations they previously have been confronted with in public transportation, as well as on technology use and acceptance (e.g. cell phone, internet usage). The questions of the pre-interviews varied slightly in detail for the three groups, but roughly stayed within the mentioned domains. For the pre-interviews with the immigrants an interpreter was present. After the pre-interview, participants received their personal DCP package. They were instructed to actively look for barriers that occur for them personally while using public transportation and further to document them. Informational background was provided, that results would be used for the conception of a mobile assistance service.

Probing. Over two weeks participants had to find as many barriers as possible by traveling through the city for their daily routines. The data acquisition took place in August 2010. Due to different personal reasons of the participants (e.g. illness) the number of active participants decreased during the data collection. After the data collection probing packages of only 9 elderly, 11 illiterates and 7 immigrants were delivered.

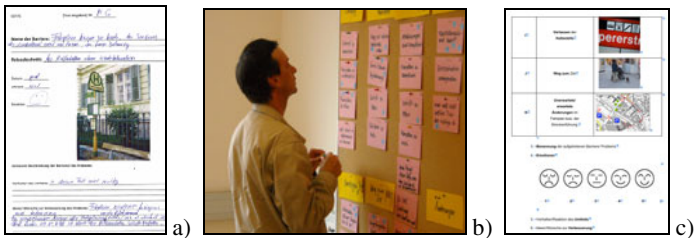


Fig. 1. a) Reprocessing in the diary, b) post focus group participant prioritising barriers, c) extract from the checksheet

Reprocessing. After two weeks of data acquisition participants returned the disposable cameras as well as the textual or oral diaries to their end user organisations. After photographs had been developed, participants glued each photograph to the corresponding diary entry (Fig. 1a) with assistance of co-workers from the end user organisations. As the immigrants were allowed to write diary entries in their mother tongue, an interpreter supported the elaboration of the immigrants' probes.

Focus Groups. In three separate focus groups participants presented their reprocessed data sheets with the barriers. Then, barriers were clustered per leg of journey (similar or identical barriers were merged), written on cards and pinned to a wall. Each participant got 20 adhesive dots to prioritise the clustered barriers per leg of journey

by sticking as many adhesive dots as they want to the cards with the barriers they think are worst (Fig. 1b). Finally, this resulted in a list of prioritised barriers per leg of journey and user group. Focus groups took 2.5 hours with 6 to 8 attendees. Not all probing participants were involved in the subsequent prioritisation process. Nevertheless, all reprocessed diary entries were included as a basis for discussion within the focus groups. Finally, each participant got financial compensation and was handed a monthly ticket for public transportation.

4 Results

Quantitative analysis of the probes showed that number of identified barriers, as well as the average number of identified barriers per person was quite variable, especially between the groups of elderly and illiterates compared to the group of immigrants. Elderly persons and illiterates detected more different barriers than immigrants (Tab. 3). The participants roughly managed to take a picture for every diary entry they made. Summing up all three groups, in 5 cases they made two pictures depicting the same barrier and twice there was a diary entry without a picture.

The number of suggestions made for every detected barrier varied as well. Elderly gave suggestions for 73% of their total number of detected barriers and illiterates for 89%. The immigrants suggested improvement for only 13% (Tab. 3).

Table 3. Number of detected barriers, diary entries, and observations per end user group

	Elderly (N=9)	Illiterates (N=11)	Immigrants (N=7)
number of different barriers detected	30	30	16
total number of diary entries	51	70	16
Average number diary entries per person	5.7	6.4	2.3
number of constructional and infrastructural barriers detected	46	69	15
number of psychosocial barriers detected	5	1	1
number of suggestions for improvement	37	62	2
amount of observed behavioural reactions in the surrounding	24, (47.1%)	29, (41.4%)	1, (12.5%)

By means of the different type of used material, participants from all user groups were able to report constructional or infrastructural issues and limitations (e.g. step entry to tramway, no rain shield at the station). Participants mostly documented barriers, which referred to their lowered ability to perceive and understand information properly and immediately.

Interestingly, although the instructed goal of the study was to detect barriers, which realistically could be avoided by the use of a continuous mobile assistance (e.g. navigator app), also psycho-social issues (e.g. unfriendly bus drivers, crowding, smell

disturbance) were noted into the diary. Especially elderly participants mentioned their perceptions in relation with social aspects and experienced discomfort (Tab. 3).

5 Discussion and Conclusion

According to the results we assume that the shaping of the packages for elderlies was well chosen for study purpose in general. Elderlies might need the possibility to reflect on additional impressions, even if they do not match the study focus, in order to keep them on track. However, due to the well-structured diary entries, it was easy to filter the non-focus entries out. Nevertheless, for further application of the DCPs, a refinement of the briefing, the checklist and a reminder task should be considered.

Overall, there were differences in the amount of detected barriers between the groups. Immigrants provided less data than the other groups. Although this user group was similarly instructed on the purpose of the study, they captured less barriers related to constructional and infrastructural issues, but also related to social interaction with others. Intercultural concerns appeared when husbands of participants forbid their wives to continue in the project. Although the procedural details were instructed in respect of any ethnic principles, people feared to violate their religious rules. Thus, in order to enhance compliance of immigrants, for further studies, we suggest to involve not only the participant himself/ herself to the study procedure, but also the surrounding family. As language barriers are persistent for immigrants, special translation and support efforts must be afforded for allowing immigrants to appropriately follow the instructions and discussions.

Although illiterates were impaired in writing and reading, they produced the largest number of diary entries in audio form. Most participants used the voice recorder without any problems, as well as they reported positive experience to document impressions without being constraint due to their minor writing skills. Another possible reason for this result is seen in the manipulation of the emphasis of verbal instruction and the related motivational effect. Illiterates were more motivated to solve the tasks if they were reinforced being ‘researchers themselves, helping scientists to see the world with their eyes’, and also having the possibility to improve life quality of others, especially their corresponding social group.

Gaver and colleagues [7] state that their ‘results are impossible to analyse or even interpret clearly, because they reflect too many layers of influence and constraint’. In comparison, data from DCPs can be processed more easily, because of minor diversity in the data. However, every researcher should always reflect about the preferred kind of data required to follow research purpose. Accordingly, DCPs should be applied to projects or studies that are interested in a specific aspect of a target group and not in their whole everyday living. We assume that DCPs can be applied to serve as a substitute of traditional methods for requirement analysis (e.g. interviews, focus-groups). They are easily adaptable to the characteristics of the target group (e.g. verbal skills), as well as they provide data that can be processed quickly.

Generally, for the application of participatory methods, tools have to be adapted to the special needs of the potential ‘researcher’ carefully, otherwise relevant and important data might get lost. In contrast to common Cultural Probes, DCPs draw a less holistic picture of a users living context. The data is directed to the exact research

question only and therefore provides more focused results. However, due to the pre- and post-data-collection contact with the user groups, it conveys a keen sense of user experience in a certain context. When Gaver and colleagues used domestic probe packages in 2004 [7], they used a purposely uncontrolled and uncontrollable approach. They provided participants with many diverse probing tools, such as a dream recorder or a friends and family map. It can be assumed that this method provides results of a wider and more global impression of how people live and how to design for them. Though, conducting Cultural Probes in terms of requirement analysis is a completely different approach. In order to extract concrete problems and user needs in a certain context, a more directed method is useful.

Altogether, we experienced the DCPs as a well performing method in terms of providing structured and pre-sorted data on the need of solutions for certain barriers. Although we do not want to claim about the novelty of our method too much, there is novelty concerning the application to this setting. The method generated useful data, which was processed in further tasks for the improvement of public transportation for people with special needs through a mobile assistance service. The raw data collection, the questions for the improvement of the barriers, the pre-interviews and subsequent focus groups including a prioritisation process, made the method expedient and valuable in terms of requirement analysis, which can be easily adapted to a different context.

References

1. Gaver, B., Dunne, T., Pacenti, E.: Design: Cultural Probes. *Interactions* 6(1), 21–29 (1999)
2. Palen, L., Saltzman, M.: Voice-Mail Diary Studies for Naturalistic Data Capture under Mobile Conditions. In: *CSCW 2002: Proc. of the 2002 ACM Conference on Computer Supported Cooperative Work*, pp. 87–95. ACM Press, New York (2002)
3. Hutchinson, H., Mackay, W., Westerlund, B., Benderson, B.B., Druin, A., Plaisant, C., Beaudouin-Lafon, M., Conversy, S., Evans, H., Hansen, H., Roussel, N., Eiderbäck, B.: Technology Probes: Inspiring Design for and with Families. In: *Proc. CHI 2003*, pp. 17–24. ACM Press, New York (2003)
4. Mättelmäki, T., Battarbee, K.: Empathy Probes. In: *Proc. PDC 2002*, pp. 266–271. CPSR (2002)
5. Hulkko, S., Mättelmäki, T., Virtanen, K., Keinonen, T.: Mobile Probes. In: *Proc. NordiCHI 2004*, pp. 43–51. ACM Press, Tampere (2004)
6. Crabtree, A., Hemmings, T., Rodden, T., Cheverst, K., Clarke, K., Dewsbury, G., Hughes, J., Rouncefield, M.: Designing With Care: Adapting Cultural Probes to Inform Design in Sensitive Settings. In: *Proc. OzCHI 2004*, pp. 4–13. Ergonomics Society of Australia, Brisbane (2004)
7. Gaver, W., Boucher, A., Pennington, S., Walker, B.: Cultural Probes and the Value of Uncertainty. *Interactions* 11(5), 53–56 (2004)

Image Retrieval with Semantic Sketches

David Engel, Christian Herdtweck, Björn Browatzki, and Cristóbal Curio

Max Planck Institute for Biological Cybernetics,
72076 Tübingen, Germany

{david.engel, christian.herdtweck, bjoern.browatzki,
cristobal.curio}@tuebingen.mpg.de

Abstract. With increasingly large image databases, searching in them becomes an ever more difficult endeavor. Consequently, there is a need for advanced tools for image retrieval in a webscale context. Searching by tags becomes intractable in such scenarios as large numbers of images will correspond to queries such as “car and house and street”. We present a novel approach that allows a user to search for images based on semantic sketches that describe the desired composition of the image. Our system operates on images with labels for a few high-level object categories, allowing us to search very fast with a minimal memory footprint. We employ a structure similar to random decision forests which avails a data-driven partitioning of the image space providing a search in logarithmic time with respect to the number of images. This makes our system applicable for large scale image search problems. We performed a user study that demonstrates the validity and usability of our approach.

Keywords: Content-Based Image Retrieval, Sketch Interface, Semantic Brushes, Real-Time Application, User-Study.

1 Introduction

There are millions of image searches performed every day which to date rely primarily on text-based queries. With the advent of increasingly powerful computer vision systems for object detection, segmentation and tracking, and the introduction of large labeled image databases such as LabelMe [1], the opportunity arises for more advanced image retrieval tools to exploit this additional information. In this paper we introduce a novel image retrieval framework for finding images based on semantic sketches in large labeled databases.

Traditionally, content-based image retrieval (CBIR) systems rely primarily on image statistics and machine learning techniques to select matching images from a database. This might not be the optimal way to approach the problem since it neglects sources of high-level information such as image annotations. Given the recent progress in computer vision it is reasonable to expect a steep rise in the number and availability of labeled images in the near future.

We propose a retrieval system that allows the user to formulate semantic queries intuitively rather than working with photometric queries. As opposed to many other sketch-based CBIR systems we do not require the user to draw detailed sketches of

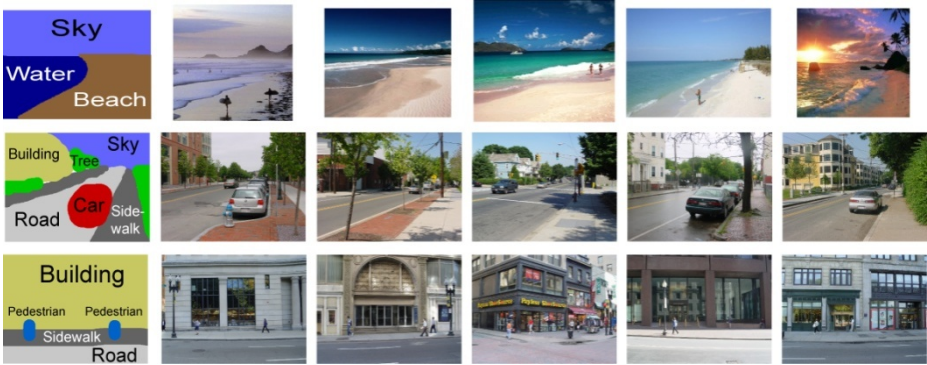


Fig. 1. Two example queries for street scene images and one for coastal images with their top five matches. Colors in the query sketch denote semantic classes. The text annotations are not part of the query and are shown here for illustration purposes only.

the objects. Our intuitive interface enables the user to indicate the semantic composition of the desired image with the help of semantic brushes (such as a brush for the classes “car” or “sky”). In such a scenario text-based searches (e.g., Google Image Search) would fail because they do not take into account the spatial relationships of the classes. Since we operate on high-level information, searches can be performed very efficiently using a tree structure, in contrast to linear methods which would be infeasible in large-scale retrieval scenarios. Our intended application is finding images that roughly match a user’s wishes, not a target search for one specific image which would require higher developed sketching abilities from the user.

To evaluate and validate our system we performed a user-study with 10 participants. They were asked to sketch street scenes and rate the images that were retrieved by our system. For the user study and the algorithmic evaluations we used the StreetScenes database [2] which contains more than 3500 images with labels of eight classes (pedestrian, car, bicycle, street, sidewalk, building, sky and tree). This database provides a suitable testbed for our algorithm as it contains a large number of images from one scene category. It can be viewed as a dense sampling of a small part of image space akin to what would result if a computer vision algorithm would automatically label a vast quantity of images.

The main contributions of our method are: the usage of high-level semantic sketches, its computational efficiency which makes it applicable to large-scale image searches, its robustness which leads to very low requirements on users sketching abilities, and its validity as demonstrated by a user study.

In Section 2 we describe previous work in this field and contrast the proposed framework to it. Section 3 describes the employed distance measures, the decision trees and the interface that comprises this system. Section 4 details the algorithmic evaluation and the user study that has been done to validate this method. Finally, Section 5 provides a summary and a discussion of future work.

2 Related Work

The need for systems that search images based on user queries has motivated a large body of research literature. In early studies the user was asked to specify the query in terms of visual features such as color or texture by drawing (query by image content, QBIC e.g. [3, 4]) or through example images (query by visual example, QBVE). The latter approach allows to calculate more complex image correlation measures (e.g. [5, 6]), yet, both approaches suffer from the so-called “semantic gap” i.e., the lack of correspondence between visual and semantic features. They yield results that have the desired low-level properties (e.g., containing a black shape) but may not fulfill the user’s semantic wishes (e.g., containing a black dog versus containing a black car). For a concise review of earlier CBIR approaches and an extensive list of references, the reader is referred to [7].

One approach to bridge this “semantic gap” is to use text-based queries as offered by image search services such as *Google*, *Bing* or *Flickr*. These approaches employ semantics in form of keywords that are assigned to images, text surrounding images in web pages, as well as manually and automatically created annotations of objects, regions and scene classes. Purely text-based systems, however, do not allow the user to specify an image composition. This can be addressed by allowing the user to draw a query image using regions of photometric patches (e.g., [8]). These patches are generated in an unsupervised fashion from training data. This is a recent approach of a query by semantic example (QBSE) technique. In systems such as [9] the user specifies an image from which a computer vision system extracts semantic properties. These properties are then compared to the images in the database to retrieve images that are semantically similar to the query image. Such systems are often not applicable for real-time searches in large databases, because of the difficult similarity judgments needed to find matches. Further reviews on semantic image retrieval can be found in [10, 11].

A quite different approach to the problem of image retrieval is photo montage or photo synthesis which aims to create the image the user has in mind instead of searching for a similar image in a database. The systems described in [12] and [13] allow the user to specify semantic regions similar to our system. Based on this input the system automatically retrieves image parts and stitches them together to form a coherent image. Sketch2Photo [14] also lets the users specify objects at any position in the image, but also requires them to sketch the objects themselves and annotate them with text labels. This gives the user the freedom to use any object label that an internet image search can reasonably retrieve images for, and also allows finer control over the objects’ appearance, but requires good sketching abilities and iterative refinement of the results.

Finally, the usability of image retrieval tools has been the topic of research. A review on semantic search tools can be found in [15], for references on image retrieval systems and their evaluation confer to [16, 17].

3 Methods

Our retrieval algorithm employs a tree structure similar to a random decision forest [18] to retrieve candidate matches from the database and a fine grained search through

the returned matches to determine a ranking. This scheme allows it to perform very fast searches (on average less than one millisecond per search in a tree containing one million images on a current office PC) with a complexity of $O(m \log n)$ where m is the number of trees and n the number of images (cf. Fig. 7).

3.1 Matching-Cost Function

Central to our image retrieval system is the definition of a cost function $C(Q, I)$ which measures the quality of the match between a query sketch Q and an image I with annotation A . This cost function allows the algorithm to rank the images and present a few top-ranking search results to the user. Desirable features of the cost function are a high correlation between the returned matches and the image the user had in mind (which will be discussed in Section 4.2), and robustness against the bad drawing skills of the average user (see Section 4.1 and Fig. 8).

We define a straight-forward cost function that can be evaluated very quickly using the scalar product between query sketch and distance transformation summed up over all classes in the image and the sketch: $C(Q, I) = \sum_i^N \langle Q_i, D_i \rangle$. An intuitive interpretation of this scheme is that we accumulate the distance that each pixel in the query image Q_i has to travel to the nearest pixel containing the respective object in the image. Using the distance transformation to calculate the cost function affords the nice properties of being intuitively plausible and making the search robust against imprecise sketching. If the annotation A of the image does not contain all objects that are present in the query Q , we add a high penalty κ to the cost function:

$$C(Q, I) = \sum_i^N \langle Q_i, D_i \rangle + \kappa(Q, A)$$

$$\kappa(Q, A) = \begin{cases} 0, & \text{if all classes from } Q \text{ exist in } A \\ \text{const}, & \text{otherwise} \end{cases}$$

A depiction of the evaluation of the cost function is shown in Fig. 2. Using such an intuitively plausible cost function allows us to easily optimize and augment our system. Linear weighting of the different object classes and adding additional cost terms to represent possible further query properties such as the color of the objects are straightforward augmentations of the cost function. Fig. 3 shows how the results are reweighted when adding a cost term for color $C_{col}(Q_{col}, I)$. We implemented this cost term by computing the earth mover distance (*EMD*) between the color histogram Q_{col} the user has specified for each region (black regions denote regions where the user does not care about the color) and the normalized histogram of the region in the image where the object is present. The *EMD* provides a suitable distance measure between two color histograms and yields better results than for example comparing the mean colors of two regions:

$$C_{col}(Q_{col}, I) = \sum_i^N EMD(Q_{col,i}, hist(A_i \times I))$$

The augmented cost function is as follows:

$$C(Q, Q_{col}, I) = \sum_i^N \alpha_i \times \langle Q_i, D_i \rangle + C_{col}(Q_{col}, I) + \kappa(Q, A)$$

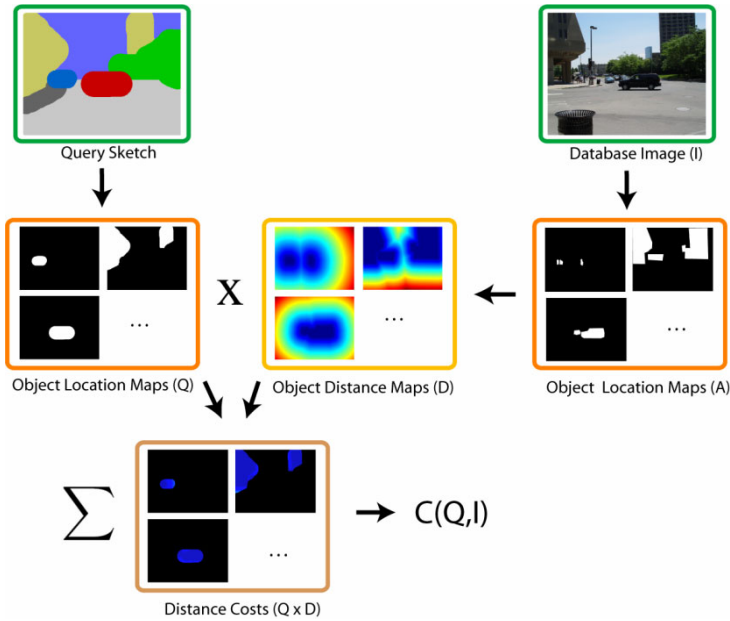


Fig. 2. Pipeline for evaluating the matching cost between an image I and a query sketch Q : for each semantic class a binary object location map is created. A distance transformation is applied. It yields a map that contains the distance from each pixel to the nearest pixel of an object. The query sketch is also translated into a set of object location maps which are multiplied pixelwise with the distance maps. Summing over all image locations and classes results in the matching cost.



Fig. 3. Constraining the image search: The top row shows a search for a car on a street in front of a building. The second row shows the results when searching for the same configuration with the additional constraint that the car should be white. The third row shows the same search, but constrained to gray buildings.

At this point the α 's are set by hand. As the main evaluation criterion for such a system is how well the results match what the user had in mind they could easily be optimized with further user studies.

3.2 Decision Trees

For small datasets that contain only a few thousand images the evaluation of the cost function for each image is feasible albeit time-consuming. As the size of the labeled dataset grows, a linear search becomes intractable especially in the context of webscale applications. However, for most such applications we do not need the full ranking of all the images in the database but only need to retrieve a couple of the top ranked images. Since the similarity of two images depends critically on the query (e.g. which image properties we are looking for) we cannot examine the similarities for each class separately but have to take the whole image into account at each decision node. This observation implies that an algorithm cannot factorize the problem, which makes it exponential in its nature.

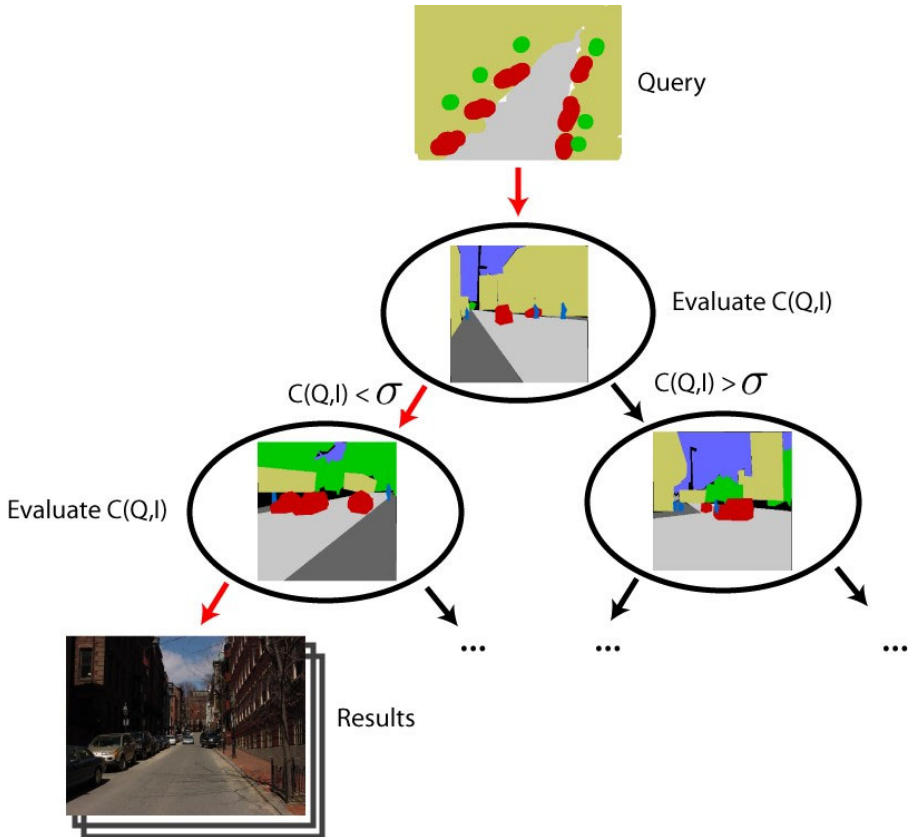


Fig. 4. Illustration of our decision tree. At each node the matching cost to a given exemplar is computed. The threshold σ determines whether the left or right child is visited next.

We address the problem of retrieving a couple of high ranking matches by using a heuristic inspired by random decision forests [18] (see Figure 4). Each random decision tree in our forest contains a pivot annotation $a \in A$ and a threshold σ . At this node the cost function $C(Q, I)$ for an incoming element is evaluated. If it is smaller than σ it is forwarded to the left child and otherwise to the right child of the node. During creation of the trees we select a random pivot element, calculate the costs to all images at that node and take the median of the costs as threshold σ . Taking the median guarantees that the resulting tree will be balanced and all searches can be performed in $O(\log n)$ time. We stop splitting the nodes when the number of elements at one node drops below a threshold (in our case five images). Consequently, any search in a tree returns one to four candidate matches, but further matches can be retrieved by traversing the tree backwards (backtracking).

By evaluating the cost function with respect to a pivot element at each node we achieve a data-driven partitioning of the image annotation space. Different pivot elements lead to different trees which highlight different aspects of the cost function. To make full use of this we search through a forest containing m trees (usually 20). We obtain the final presentation order by ranking all returned results according to the cost function. Consequently, the total runtime of a search is $O(m \log n + m k)$ where k is the time it takes to evaluate the cost function for the resulting matches returned by one tree.

We show that the results returned by our forest approximate the matches returned by a linear search in Section 4.1. We further demonstrate that our distance measure is highly correlated with subject ratings of the user study described in Section 4.2. Together, these results show that our approach is a valid scheme to quickly retrieve images from a database based on semantic sketches.

3.3 The Sketch Search Interface

We created a painting tool that enables the user to specify the composition of desired images (cf. Fig. 5). The tool offers a collection of semantic brushes that allow the user to specify object locations. These brushes can represent objects such as cars or bicycles but also image regions such as sky or road. To distinguish different brushes we assign a unique color to each of them. The color value itself is not relevant, since color is only used to denote the type and location of an object, not its appearance. Which brushes are available depends on the labels found in the current image database. For our evaluations we used the eight classes contained in the StreetScenes database (see Section 4).

To create a query image the user selects scene elements and draws them onto a canvas. The handling of the tool resembles that of common image editing programs. This way of composing a scene is intuitive and self-explanatory, as confirmed by the participants in our user study (see Section 4.2).

To specify an object it is not necessary to draw its precise shape. It is sufficient to mark the region in the image where instances of its class should occur. By drawing two cars, as shown in the example sketch in Fig. 5, the user specifies that the two areas marked as “car” (red) should contain objects labeled as “car”. This constraint is fulfilled as long as there is at least one matching object in each of those regions, thus allowing the presence of cars in other regions.

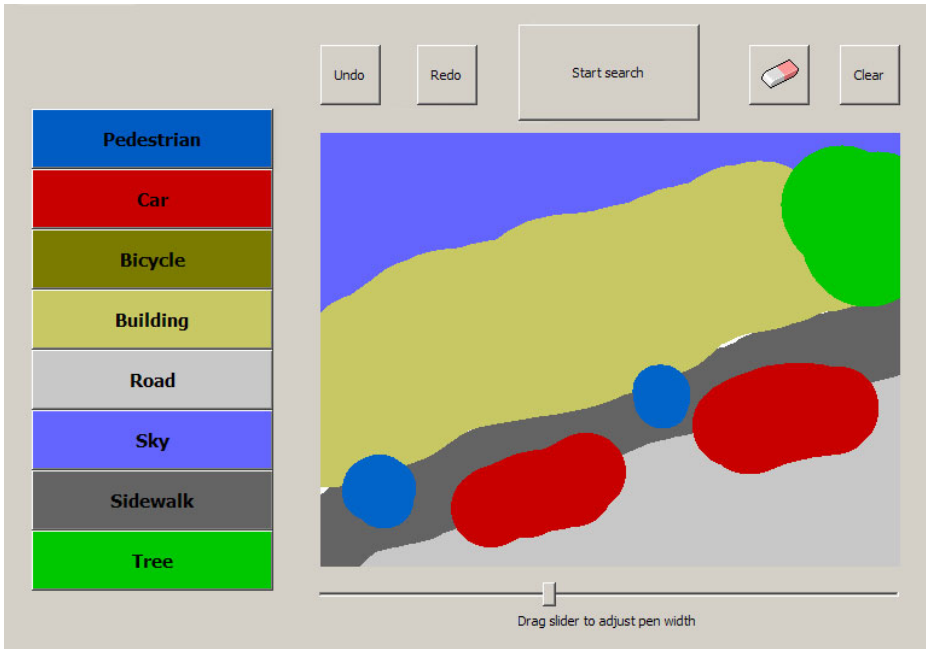


Fig. 5. Sketch Search user interface. The user chooses objects to include in the scene from the object palette on the left. The canvas depicts a drawn street scene that is used as query for the image search. It shows a street scene containing a road, a sidewalk, two cars, two pedestrians, buildings, trees and sky.

4 Evaluation

In this Section we present the evaluation of our system. We evaluate and validate our method using the StreetScenes database [2] which contains 3547 images taken in Boston together with annotations of eight major classes: Pedestrian, Bicycle, Car, Street, Sidewalk, Building (including stores), Sky and Trees. Note that our algorithm does not distinguish objects (e.g. cars, people) and semantic regions (e.g. sky, street) and deals with them quite naturally. These classes are well suited for our envisioned application since they could potentially be labeled automatically by a computer vision algorithm. The number of categories might seem small when compared with the much higher number of classes found for example in the LabelMe database. But, most of the classes from LabelMe are irrelevant for the proposed large-scale image retrieval task as they have too few occurrences in the database.

Obtaining human annotations is expensive and time consuming which implies that our algorithm would mainly operate on computer generated labelings. These annotations would only contain labels for categories that are accessible to vision algorithms, which amounts to a few high-level categories for the near future. Gender specific searches are for example not plausible since algorithmic differentiation between male and female persons is a difficult task. Lastly, it has to be mentioned that there is a strong correlation between classes naturally occurring in images (e.g. “cars”

and “road” often appear together in an image while “car” and “table” do not). This further reduces the number of classes that have to be present in a tree.

4.1 Algorithmic Evaluation

As first evaluation we perform automated searches using queries formed from ground truth annotations from the database. We plotted the results produced by our algorithm and the results obtained when evaluating the cost function for each image individually in Fig. 6. The results show that our algorithm is able to approximate the linear search through the database with a logarithmic search in our random decision forest.

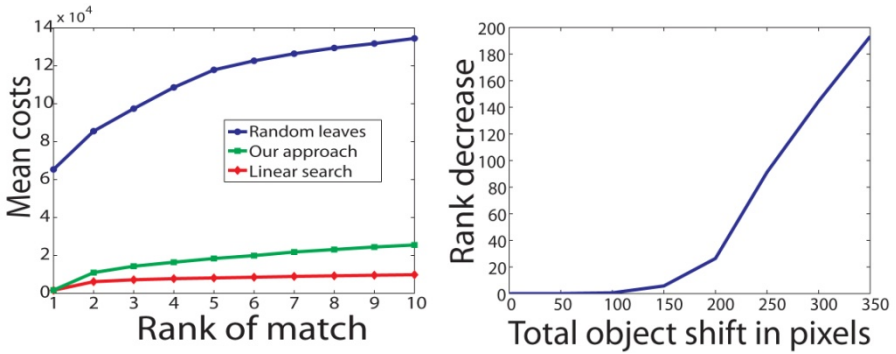


Fig. 6. Left: Comparison between results for our approach, a linear search through all images and a ranking of random tree leaves according to our cost function. The graph shows the mean costs associated with the top 10 matches. **Right:** The figure shows the decrease in retrieval performance for shifts of object location. The average decrease of matching rank compared to the unshifted queries is plotted on the y-axis.

Insensitivity to variation in object positions in query images is crucial to provide a level of robustness that is needed to deal with the low fidelity of user sketches (see Section 4.3). We evaluated the robustness of our system by shifting the positions of object classes in horizontal or vertical direction by a random amount. However, we set the total number of pixels that all object classes were shifted in one image to a fixed value. In this case a shift by 320 pixels means that each of the eight object classes we use is shifted by 40 pixels on average. Fig. 6 shows the decrease in retrieval performance with respect to the original queries. For smaller shifts up to approximately 150 pixels the results differ only slightly from those obtained by the original queries. It is important to note that query images had a dimension of 320x320 pixels. Shifting an object class by a large distance can result in its complete removal from the scene, contributing to the steeper performance decrease at summed shifts of more than 200 pixels (which can be expected to be intended by the user). The evaluation shows that our algorithm is robust against variations in the sketched object location up to a certain degree.

For eight classes and a 32x32 descriptor each image can be represented using 8096 Byte. This means that a million images would take up 8GB of memory. This memory problem is present for every image retrieval system as they need to keep the image

descriptors in memory during runtime. Fortunately, due to the linear nature of the distance transformation the dimensionality can be greatly reduced using a principle component analysis (PCA). Using PCA we can encode 99% of the variance of the descriptors with just 34 dimensions for this highly redundant data-set. When stored in a single precision float vector the compressed descriptor takes up only 136 Byte. The evaluation of the cost function is then preceded by a matrix multiplication and addition of the means to project the principle components scores of the descriptors back into the image space. This has to be done only $\log n$ times for each tree, thus, not influencing the runtime critically. This effectively removes the memory problem, making our algorithm applicable for large scale image search problems.

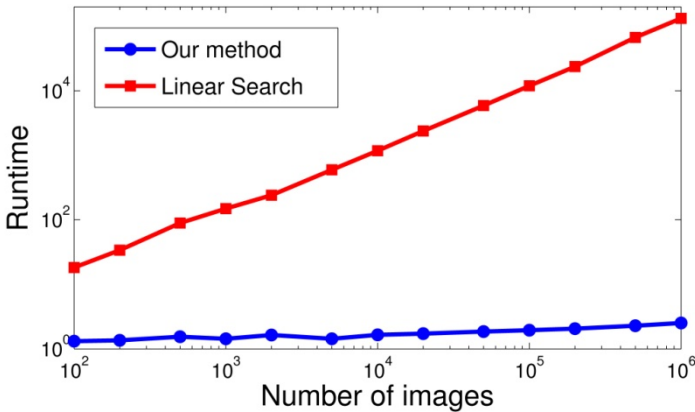


Fig. 7. Comparison between searches using our method and a linear search in databases of different sizes. Graph axes are logarithmic.

To evaluate the speed of our search we built trees for a more realistic search setting. By mirroring, shifting and subwindowing random images from the StreetScenes database we created a dataset containing one million images. All timing studies were conducted on a regular office PC with a 3GHz dual core processor and 2GB of RAM using our MATLABTM implementation of the search algorithm. Creating a tree for this dataset takes on average 3 minutes. Queries in this tree take on average 0.23 milliseconds. A linear search through all image descriptors on the other hand takes 17.4 seconds on average making it unsuitable for any large scale application. For more details on computational efficiency see Fig. 7.

4.2 User Study

In this section we show that our approach retrieves images that are not only good in an algorithmic sense but also in a subjective sense which is of key importance for a retrieval system. We conducted a user study with ten participants (7 female, 3 male, mean age was 25.6). On average each participant did 45.7 trials in 60 minutes. The drawing phase took 53 seconds on average. To ensure that participants would sketch different images in each trial they were shown an “inspiration” image taken randomly

from the StreetScenes database for one second before the drawing phase started. The subjects were explicitly instructed to take this image solely as an inspiration and not to search for this image. In the drawing phase users then sketched an image using the tool described in Section 3.3. The resulting sketch served as input for our algorithm which returned images using a forest consisting of 20 trees.

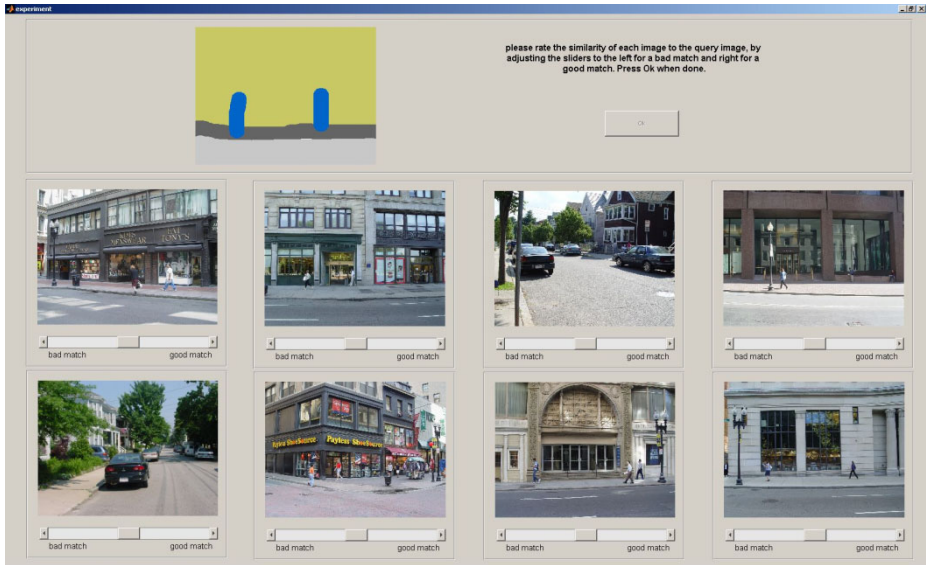


Fig. 8. Example query and resulting images in the GUI from our user study. The third image in the first row and the first image in the second are random images, our algorithm’s best matches are the second image in the first row and the last in the second row.

For the evaluation phase we selected the four best matches retrieved by our algorithm and the worst and median matches retrieved by our algorithm as well as two random images from the database. We presented the images simultaneously at randomized positions in the GUI shown in Fig. 8. Note that “worst” and “median” here does not refer to the worst/median match in the whole database but just in the subset (on average 54 images) of potential matches returned by the search. Participants were then asked to rate the similarity of the images to their sketch on a scale of 1 (bad match) to 7 (good match). Participants were explicitly instructed not to rate the similarity to the original “inspirational” image which might have created confounds when users remember only certain details about the images.

For data analysis, ratings were normalized to a mean of zero and standard deviation of one. We then calculated the average participant rating for each of the seven classes of shown images (best match, second/third/fourth best match, mean match, worst match, random image). The results are visualized in Fig. 9. A post-hoc Scheffé test (significance level $\alpha=1\%$) between the means confirms that participants gave significantly higher ratings to images that our algorithm considered to be good matches than to random images or bad matches. Furthermore, the average rating for the best match is significantly better than the rest while the second, third and fourth

matches show no significant differences. The mean rating for a “median” match is significantly worse than the top-matches and significantly better than the worst and random matches. This is to be expected as the worst of the retrieved images often did not contain all classes from the sketch.

We further analyzed directly the relation between the algorithm's cost function and the participants' ratings. In Fig. 9 we visualize the mean costs over all query results compared to the average participant ratings for them. The y-axis of the plot is normalized to one standard deviation of all participants' responses. There is a strong correlation, which is partly due to the cost term κ which penalizes the absence of objects in bad matches.

In summary, these results confirm that our algorithmic definition of a good match coincides with the human intuition, allowing our system to yield user-intended results.

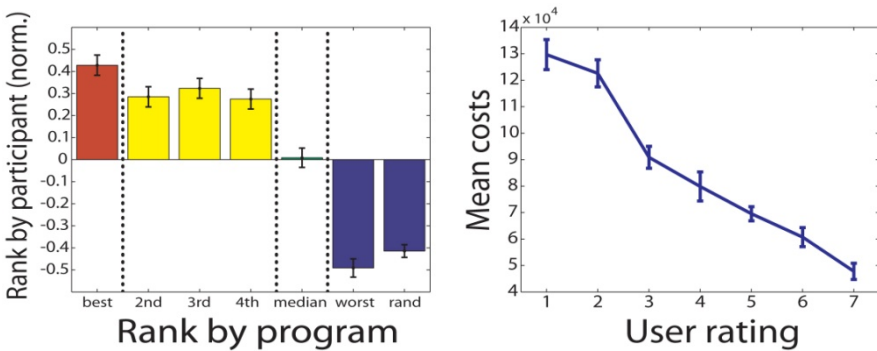


Fig. 9. Left: Normalized ratings of participants for our algorithm's top four matches, the “median” and “worst” match (see text) and two random images. Colors indicate a grouping of bars: bar heights in different groups are significantly different while bar heights in the same group are not. Right: The average costs that our algorithm assigns to potential matches show a strong correlation with participant ratings for these matches (1=bad, 7=good match). In both plots errorbars denote standard error.

4.3 Cognitive Constraints

Our user-study showed that the manual generative abilities of participants are considerably worse than their visual discriminative abilities. Fig. 10 shows some examples of images and corresponding sketches users have drawn during the study. As people are able to copy an image shown during the sketching phase, the problem seems to occur when users have to create a 2D representation of a 3D scene they have in mind. Similar observations on participants failing to reproduce the correct perspective of a scene from memory have also been shown in psychophysical studies such as [19]. This poses a general problem to most systems that rely on sketch based interfaces. Such systems are expected to produce an image matching the one the user has in mind based solely on a potentially inaccurate sketch. Our system addresses this problem by using only rough semantic sketches and a distance transformation, both of which help to suppress errors users make during sketching.



Fig. 10. Failure case: “inspirational” image at the top left and sketch created from it in the user study on the bottom left; on the right the four best matches. For some of the bad user sketches (e.g., with impossible perspective as here), no reasonable matches can be found in the database, resulting in relatively bad results.

We feel that our interface that requires only rough sketches is already at the upper bound of what can be expected from average users, especially when looking at the errors in perspective that the users make during drawing. More complex sketching systems for purposes such as “Sketch2Photo” [13] allow users to create visually pleasing images with arbitrary scene compositions. However, they only work if the user is able to produce a sensible perspective composition otherwise they can only create an unrealistic image. Our system cannot generate images de-novo but is able to find sensible matches very efficiently and robustly which seems to be a better approach.

5 Conclusions

In this paper we have presented a novel system for content-based image retrieval using semantic sketches. By utilizing a fast straight-forward cost function together with a decision tree with guaranteed depth of $O(\log n)$ we are able to provide a very fast tool. Our system yields comparable results to a linear search thus being applicable for large scale image databases. In our user-study, we have demonstrated that our system and the cost function is usable and return images matching what the user drew. These properties demonstrate that our system is useful for image retrieval in large labeled databases.

In the future we plan to do a larger scale user study in order to optimize the parameters of the cost function and accommodate user sketching behavior. Furthermore, we will apply this retrieval system in tandem with a computer vision system that provides a coarse labeling in order to provide a sketch based search tool that operates on a database of millions of images.

References

- [1] Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77(1-3), 157–173 (2008)
- [2] Bileschi, S.M.: *Streetscenes: towards scene understanding in still images*, Ph.D. dissertation, Massachusetts Institute of Technology (2006)
- [3] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: The qbic system. *Computer* 28(9), 23–32 (1995)
- [4] Jacobs, C.E., Finkelstein, A., Salesin, D.H.: Fast multiresolution image querying. In: *Computer Graphics and Interactive Techniques*, pp. 277–286 (1995)
- [5] Marée, R., Geurts, P., Wehenkel, L.: Content-based image retrieval by indexing random subwindows with randomized trees. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) *ACCV 2007, Part II. LNCS*, vol. 4844, pp. 611–620. Springer, Heidelberg (2007)
- [6] Hirata, K., Kato, T.: Query by visual example - content based image retrieval. In: Pirotte, A., Delobel, C., Gottlob, G. (eds.) *EDBT 1992. LNCS*, vol. 580, pp. 56–71. Springer, Heidelberg (1992)
- [7] Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence*, 1349–1380 (2000)
- [8] Fauqueur, J., Boujemaa, N.: Logical query composition from local visual feature thesaurus. In: *Content-Based Multimedia Indexing* (2003)
- [9] Rasiwasia, N., Moreno, P.L., Vasconcelos, N.: Bridging the gap: Query by semantic example. *IEEE Transactions on Multimedia* 9(5), 923–938 (2007)
- [10] Datta, R., Li, J., Wang, J.Z.: Content-based image retrieval: approaches and trends of the new age. In: *Workshop on Multimedia Information Retrieval*, pp. 253–262 (2005)
- [11] Liu, Y., Zhang, D., Lu, G., Ma, W.-Y.: A survey of content-based image retrieval with high-level semantics. *Pattern Recognition* 40(1), 262–282 (2007)
- [12] Johnson, M., Brostow, G., Shotton, J., Arandjelovic, O., Kwatra, V., Cipolla, R.: Semantic photo synthesis. *Computer Graphics Forum* 25(3), 407–413 (2006)
- [13] Diakopoulos, N., Essa, I., Jain, R.: Content based image synthesis. In: *International Conference on Image and Video Retrieval*, pp. 299–307 (2004)
- [14] Chen, T., Cheng, M.-M., Tan, P., Shamir, A., Hu, S.-M.: Sketch2photo: Internet image montage. *ACM Transactions on Graphics* 28(5), 1–10 (2009)
- [15] Uren, V., Lei, Y., Lopez, V., Liu, H., Motta, E., Giordano, M.: The usability of semantic search tools: a review. *The Knowledge Engineering Review* 22(04), 361–377 (2007)
- [16] Shirahatti, N.V., Barnard, K.: Evaluating image retrieval. In: *Computer Vision and Pattern Recognition*, vol. 1, pp. 955–961 (2005)
- [17] Vogel, J., Schiele, B.: On performance characterization and optimization for image retrieval. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) *ECCV 2002. LNCS*, vol. 2353, pp. 49–63. Springer, Heidelberg (2002)
- [18] Ho, T.K.: Random decision forests. In: *Conference on Document Analysis and Recognition*, p. 278 (1995)
- [19] Previc, F.H., Intraub, H.: Vertical biases in scene memory. *Neuropsychologia* 35(12), 1513–1517 (1997)

Mixer: Mixed-Initiative Data Retrieval and Integration by Example

Steven Gardiner, Anthony Tomasic, John Zimmerman, Rafae Aziz,
and Kathryn Rivard

Carnegie Mellon School of Computer Science, 5000 Forbes Ave Pittsburgh PA 15213
{sgardine,tomasic,johnz,raziz,krivard}@cs.cmu.edu

Abstract. Office administrators are frequently asked to create ad hoc reports based on web accessible data. The web contains the desired data but does not allow efficient access in the way the administrator needs, prompting a tedious and labor-intensive task of retrieving and integrating the required data. Mixer is a programming-by-demonstration (PBD) tool empowering administrators to construct ad hoc reports from diverse web sources without tedious piecemeal labor. Mixer's design builds on the exploration into end user conceptualization of data retrieval tasks from our previous Wizard-of-Oz study [39], and incorporates insights from mixed-initiative researchers into collaboration between end users and software agents. This paper justifies the design decisions that drive Mixer, focusing on general lessons for designers of programming-by-demonstration systems targeting nonprogrammers. We evaluate Mixer by performing a user study showing that administrators are able to accomplish programming tasks without needing to understand programming concepts for data retrieval and integration.

Keywords: programming by demonstration, end user programming, mixed initiative, data integration.

1 Introduction

As the size and richness of the web has steadily increased over the last decade, users have ratcheted up their expectations for the scope of information easily available from the web. Web interfaces, in contrast, usually permit access to the information they expose in a highly constrained manner. The gap between the form of the information required and the form provided by the deployed interfaces is bridged in practice by human intelligence. In particular, office workers in an administrative capacity are regularly assigned mundane, repetitive data integration tasks, entailing the gathering of information from several sources into an *ad hoc report*, often in response to an email [34]. Administrators do not consider these tasks difficult, but they do consider them very tedious. The repetitive and procedural structure of the tasks makes them ripe for automation; however, the actions taken in response to the retrieved information generally require human judgment. This combination of automation and human judgment invites a mixed-initiative approach that weds the administrator's understanding of the desired report with a programmatic agent actually performing the bulk of the mundane retrieval.

As an example, suppose a university dean wishes to investigate previous collaborations between her university and a certain research lab, and further that she has assigned her assistant the task of finding all professors in the university who have published a paper with someone from the lab. The straightforward solution is to look up in turn each professor's publications in a digital library and store all of their collaborators from the lab in some intermediate location, such as a column of a spreadsheet. This solution, while effective, illustrates well the tedium involved, as it requires the administrator to perform the same series of clicks and copies and pastes, each time with different input, until the output report is complete.

The tedious, repetitive nature of the tasks evokes the concept of *programming by demonstration* (PBD). The application of PBD to administrative data integration tasks follows from the insight that once the user has shown how to look up a single example, the system has sufficient information about the procedure to look up the rest of the examples. In an effort to realize the promise of PBD in facilitating administrative data integration tasks, we developed Mixer, a Mixed-Initiative PBD system that allows users to train an agent to perform the tasks. Our current implementation builds on our previous Wizard-of-Oz study, which demonstrates the effectiveness of a spreadsheet-like user-created form as a medium of communication between a human user and a simulated computer agent. Users were quite successful in using the mocked-up system, but they struggled with the following issues: (1) specifying 1-to-many relationships in a manner useable by the agent, (2) specifying precision in the retrieved report, and (3) selecting meaningful segments of text on the page. Mixer as presented here addresses these shortcomings, and also incorporates insights from other explorations of web PBD [20, 22]. Mixer presents several innovations over previous approaches. First, Mixer presents a unified modeless interface for integrating data, whether that data come from one or several data sources. Additionally, Mixer leverages the insights of Mixed-Initiative design to facilitate collaboration between the user and the agent to accomplish the user's goal.

To evaluate our design decisions, we conducted an evaluation with real administrators. The administrators were asked to retrieve multiple items from a single data source and to link information across multiple data sources. The evaluation results show that: (i) Using the Mixer table based interface, administrators can conceive of, create, and use forms that effectively communicate to the Mixer agent both the information they want and the information the agent needs to automate the task; and (ii) administrators recognized the value of automating this type of mundane task and indicated they would incorporate a mixed-initiative tool like Mixer into their work practices.

The remainder of this paper is organized as follows. First, we describe the design of Mixer. We then describe the study performed and the results obtained. Next, we discuss the implications of the present study. Lastly, we situate this work within the related work in the literature, and conclude.

2 Design

At a high-level, interaction with Mixer requires the user to construct the first row of a table, and in doing so, the user demonstrates to the agent what information they want and where this information can be found. When this row is complete, the user releases

the agent to follow the pattern until the retrieval is complete. If the user desires a subset of this information, they can export the resulting table to a spreadsheet and use the spreadsheet to make the subset they desire selectable. Below we detail an example, depicted in Figure 1, of how the interaction works. In the example, the user looks up the names and affiliations of the coauthors of a particular researcher using an online interface which allows accessing this information for one publication at a time.

1. Orange highlight indicate what item can be selected.

2. selected item fills in first column.

3. Selected item fills in first cell of second column. Mixer adds coauthors in subsequent cells.

4. Select "fill table" button to execute Mixer program.

5. Select "export sheet" button to export data to a spreadsheet program.

paper title	author name	author affiliation	fill table
Feeding in human-robot con...	Biige Matlu	Carnegie Mellon University,	cancel
	Toshiyuki Shima	ATR, Kyoto, Japan	cancel
	Takayuki Kanda	ATR, Kyoto, Japan	cancel
	Hirosi Ishiguro	Osaka University, Osaka, Jap	cancel
Nonverbal leakage in robots			
Designing gaze behavior for			
The design of gaze behavior f			
Robots in organizations: the			

paper title	author name	author affiliation	fill table
Feeding in human-robot con...	Biige Matlu	Carnegie Mellon University,	cancel
	Toshiyuki Shima	ATR, Kyoto, Japan	cancel
	Takayuki Kanda	ATR, Kyoto, Japan	cancel
	Hirosi Hagita	ATR, Kyoto, Japan	cancel
Nonverbal leakage in robots	Fumihiko Yamanka	ATR, Kyoto, Japan	cancel
	Takayuki Kanda	ATR, Kyoto, Japan	cancel
	Hirosi Ishiguro	Osaka University, Osaka, Jap	cancel
	Hirosi Hagita	ATR, Kyoto, Japan	cancel
Designing gaze behavior for	Jodi Forlizzi	Carnegie Mellon University	cancel
	Hirosi Hagita	Carnegie Mellon University	cancel
The design of gaze behavior	Biige Matlu	Carnegie Mellon University	cancel

Fig. 1. A user interacting with Mixer to extract the coauthors of a researcher based on the researcher’s papers on the ACM Digital Library

Users begin using Mixer by navigating their browser to the first page they wish to retrieve information from (referred to as the target page). In this example the page contains a listing of the publications of the researcher, with a brief description of each publication and a link to a detailed record of the publication. Once there, the user clicks on the Mixer button appearing within the browser's chrome. This click causes two actions. First, the Mixer workspace (referred to as workspace) appears in a frame to the right of the target page. Second, Mixer augments the target page, highlighting any element the agent can accept as an input in orange.

To add an element to the workspace, the user right-clicks the element and selects "Copy to Mixer" in a context menu. In response, Mixer constructs a new column in the workspace, gives it a heading, fills in the first row with the element the user selected, and fills in the remaining rows with all elements that match the user's selection (Figure 1A). In this case the user selects the title of the first publication on the list and the agent adds the title "paper title", adds the user-selected publication, and adds all of the remaining publications to the column. The cell at the top of the filled-in column has a white background, indicating a human selection while the cells below have an orange background, indicating that the agent selected these elements. In this example, the user only needs the publication title from the first page; however, if the user required additional elements from this page, right-clicking and selecting "Copy to Mixer" would cause the agent to create additional columns. Earlier Mixer interface designs allowed users to simply copy and paste elements from the target page to the workspace. However, we observed that people had trouble understanding what was "legally" selectable since they could never see the underlying data scheme. In addition, they had trouble copying and pasting text that appeared as a link. In the current design, the highlighted elements on the target page are intended to clarify what can be selected and the right-click action allows users to add an element that is a link, without navigating away from the target page.

In our previously reported evaluation of the Mixer interaction design, we noted that some participants struggled to create an effective table. Using the example of the publications, many would copy and paste the first publication title into the top of the first column and then they would copy and paste the second title on the list to the top of the second column. To prevent users from making this mistake, Mixer automatically fills in the first column as soon as the first element is selected.

Now that the user has a column with all of the publication titles, the user next needs to demonstrate that they also want authors to go with each publication. To advance the task, the user navigates through the publication link causing the target page to change from the publication listing to the publication detail page.

Mixer detects that the user's action depends on the contents of a cell in the workspace, i.e. the link corresponds to the first publication. Accordingly, Mixer begins an implicit loop over all publications in the column. An additional wrinkle which does not appear in the demonstrated task, is when a user must type input from the workspace into a query form on a separate page. In such cases, Mixer prevents direct typing because the agent needs to be shown an explicit connection between the contents of a cell in the workspace and the input to a query. When a cell's contents are copied to the clipboard and then pasted into a form's widget, Mixer is able to deduce the connection. When the cell's contents are simply typed into the widget, however, Mixer cannot be certain that the query input in fact comes from the

workspace entry rather than from elsewhere in the page; to avoid this problem, Mixer issues a warning to the user when directly typing into a query form. In a more extreme case, the administrator might modify the contents (e.g. stripping off the first name and using only the last name), making the matching of the contents to the workspace quite difficult. This tension between re-using a variable by value and by reference has a long history in PBD, see e.g. the discussion of distinguishing constants from variables in Myers [26].

Mixer augments the publication detail page: selectable elements are highlighted in orange. These elements include the publication's authors' names, as well as the venue where the publication appeared (Figure 1B). The user right-clicks the author's name and selects "Add to Mixer." In response, Mixer creates a new column, adds the title "author name", fills in the first row with the selected author name, fills in subsequent cells with additional coauthor names available from the current page, and fills in the remaining rows with a dashed line, indicating that the agent thinks the user wants this information for subsequent publications. Similarly, the user adds "author affiliation" information to the table from the same page.

To release the agent to complete the table, the administrator clicks the "Fill Table" button to the right of the last column. In response, Mixer infers and executes a program. In this example the program contains a loop over all publications in the publication listing. Mixer iterates over each publication, one by one. For each publication, Mixer navigates through the link using exactly the same action sequence which the administrator demonstrated, modified only to correspond to the present publication. As each detail page is visited and its result integrated, the browser view shows exactly what is happening, giving the administrator confidence that the result is the same as if the task were performed manually. When the table is complete, Mixer plays a chime, letting the user know the agent is finished (Figure 1C).

Administrators frequently want to collect information about a subset of items that meet some criterion, for example the students who were currently failing a particular course. Since administrators are familiar with spreadsheets, we postulated that they would be able to conceptualize the filtering task as composed of the subtasks of retrieving the desired information about the whole group, then sorting on the selection attribute and cutting out all nonqualifying rows. This functionality is present and familiar in modern spreadsheets, so Mixer does not re-implement it, but rather expects the user to use a separate spreadsheet tool (our experiments used Google Docs).

Our previous simulation of the Mixer interface illuminated the way for the present implementation; however, as noted by Sundström et al [32], the Wizard-of-Oz methodology only allowed a certain level of familiarity with the algorithmic material. The actual implementation explored some new design potential and constraints. One notable example is that the implementation must take into consideration the time taken for the actual retrieval, balancing the time consumed by the network latency inherent in retrieving information from the web with the user's valuable time and attention. The implemented system, unlike the Wizard-of-Oz mockup, does not begin with access to the retrieved data. Instead the user must wait while Mixer replays the demonstrated actions necessary for retrieval. The present design replays the actions in front of the user's eyes; this furthers communication between the user and the agent in that the user understands what is happening and why it is taking time.

Relatedly, the present design fills in as much of the table as it can as soon as the user selects a piece of data for inclusion. This refinement allows Mixer to communicate more effectively its understanding of the table as the table is constructed. We thereby lessen some of the issues participants encountered with our previous design, as to how to construct the table.

Additionally, constructing the table as soon as data is available allows more efficient use of the user's time. Specifically, in the previous design users engaged a tool we called the "resolver" to help specify if they wanted a single element, a subset of element, or all elements within a column. In an actually executing system, however, this means the user must make resolver decisions periodically throughout the retrieval process, with lags of unknown length between decisions. In addition, in designing the resolver tool we struggled to find a way for users to precisely specify what they specifically wanted that did not feel like programming. The current design takes a different tack. It drives users to collect a complete set of data, and then allows them to export the table they have made to a spreadsheet, where they can use the familiar tools of spreadsheets to sort and perform calculations. This design choice gains significant ease of use at the cost of making precision more laborious.

Two other changes from the previous design bear mention. First, whereas the previous design presented the user with the target page and invited her to select any page element, the current design instead pre-highlights the selectable elements on the page. This clearly communicates to the user which actions Mixer will understand, but decreases the ability of a user to apply Mixer to novel pages. Second, we constrain the user to copy and paste data from the workspace into a query page, rather than typing.

2.1 Wrappers

Mixer augments the target page with orange highlights to indicate selectable elements. The agent performs this augmentation by applying a wrapper to the page. A wrapper is a small piece of code that identifies the types and location of data within the page. Each time a new web page is visited, Mixer consults a database of wrappers, selects the most appropriate wrapper for the page, and applies the wrapper to the page. Based on the application of the wrapper to the page, Mixer highlights the wrapped data on the page, as shown in Figure 1, and adds the appropriate user interaction bindings. The use of wrappers represents a departure from the previous Mixer interface. The wrapper directly encodes the hierarchical relationship among the potential columns, allowing the agent to automatically fill in the first column as soon as the first element has been selected.

An obvious bottleneck for the applicability of Mixer is the coverage of its database of wrappers, raising the question of how this database would be populated. The present work makes the simplifying assumption that the database is pre-populated with all necessary wrappers. For real world usage, several possibilities exist. First, IT shops who are interested in offering Mixer capability to their customers might code wrappers for their internal websites and distribute a version of Mixer with access to those wrappers. Alternately, a crowd of third-party developers might contribute wrappers for web pages to an internet-wide repository for use internet-wide, as is done with public GreaseMonkey [6] or CoScripter [20] scripts. An extension of this approach is to deploy tools (e.g. Mash Maker [13] or reform [35]) enabling end users

to participate in this crowd-sourcing construction of the wrapper repository. Mash Maker [13] maintains exactly such a repository of wrappers contributed by end users. The Accessibilities Commons [17] maintains a similar database of web accessibility enhancements, designed to crowd-source a solution to the accessibility problem.

More formally, a wrapper overlays a relation, i.e. a list of one or more tuples, over the page. The wrapper specifies the type of the tuple as an unordered list of text fields and subrelations (with tuples specified in the same way¹). Each tuple and "leaf" field is located with an XPath expression [7], interpreted relative to its parent tuples; thus the addressing is similar to that in the Accessibilities Commons [17]. To minimally overcome the limitations of identifying fields with DOM elements (see e.g. Dontcheva et al [12]), fields may optionally refer to a regular expression matching some of the tokens within the chosen element. Thus the Mixer wrapper formalism is sufficiently expressive to wrap pages with a uniformly repeating tuple-type where tuples and fields are contiguous within DOM elements. Partly as a tradeoff for this expressivity, Mixer wrappers are nontrivial to specify; thus the sheer number of wrappers required for widespread Mixer usage seems to present a more limiting bottleneck than the limitations of the expressivity of the wrapper formalism.

2.2 Mixer Program Induction

Generally, *program induction* is the task of constructing a program based on a small number of example executions of the program. Mixer performs program induction by observing the browser actions as the user performs the demonstration, and forming a program which will be executed when the user asks to complete the table.

In order to reduce the prohibitively large search space [19] of programs consistent with the demonstration, Mixer leverages relatively strong assumptions about the Mixer domain to decompose the induction problem into several smaller sub-problems². The assumptions about the domain are:

- Universality: All loops are repeated over all instances in their scope
- Quantification: All loops are scoped over the values in a workspace column
- Task Focus: Only actions which change the workspace can affect the program

The learning mechanism records all browser actions in a log of actions. The log contains the user's actions without modification or generalization of any kind. Parallel to the log, Mixer constructs the program, which represents the best guess about the repeatable procedure whose output the user desires. The instruction set contained in the program are the same as the browser actions contained in the log, with the addition of parameterized versions of each action as well as a looping *foreach* structure. Additionally, a program begins with a preamble which loads the starting page where the replayed browsing actions will commence.

The learning mechanism appends each observed browser action to an accumulated list of steps. The first sub-problem is to classify a given sequence of steps as to whether it forms a complete unit of action, which should affect the program in some

¹ The allowance for subrelations expands the expressivity of Mixer wrappers to nested tables, also known as non-first normal form relations (see e.g. [38]).

² This decomposition was chosen to be amenable to Machine Learning classification.

way. The Mixer implementation presented here uses a simple heuristic, considering only step sequences terminated by a step inserting new content into the workspace.

Given that a unit does affect the program, Mixer next decides how it does so. The sequence of steps in the unit is decomposed into the *transition*, the *query*, and the *selection*. The query determines how data from previous columns drives widgets to perform the web lookup, and the selection selects information to be inserted into the workspace; the transition contains any preparatory steps which are not data-dependent³. The steps in the transition are simply appended to the program at the current insertion point, and afterward a new loop is appended, quantified over the entity type used by the query. The steps of the query are parameterized to depend on data from columns in the workspace, then appended inside the new loop. Finally, the insertion point is advanced to the current end of the loop. The current implementation of Mixer makes no effort to track previous positions of the insertion point; as a consequence the “undo” operation only functions back to the insertion point. Beyond the insertion point, the user cannot effect any changes to the underlying program, so in case of error must restart the session.

The selection is converted into a single atomic command for extracting all content from the visited page. While inserting new content into the workspace always appears to the user to change the workspace, only the first piece of data taken from a visited page actually affects the underlying program. The workspace, meanwhile, contains all the data from the page but only displays the pieces the user has demonstrated so far. Upon demonstration of the addition of subsequent pieces of information to the workspace, the program is unchanged; only the visibility flag for the affected column is toggled. This low-level distinction is made invisible to the user.

3 Evaluation

We recruited administrators to perform a user study aimed at substantiating the following hypotheses:

- H1: Mixer’s table-based workspace interface provides an effective method of communication between the human and the agent for data integration tasks:
 - Administrators can conceive of and express information demands through designing and demonstrating the form of the information in the workspace.
 - Administrators can make sense of, and work with, information retrieved in collaboration with an agent and presented in the workspace.
- H2: Administrators will recognize the benefit of automated data integration and would be interested in using this interface for their work.

In order to test these hypotheses we culled administrative tasks from the suggestions of participants in our previous Wizard-of-Oz study of the Mixer

³ The transition would contain interactions with widgets of a form which are universal to the tuple, but not to any particular value of a field. For example, checking a checkbox choosing to search for people rather than, say, departments, would be included in the transition; the query would continue by specifying which person to search for.

interface [39]. To accommodate privacy concerns, we shifted the tasks to different real-world domains, where we selected isomorphic tasks which pilot participants demonstrated could realistically complete within a 90 minute experiment. Because Mixer is not intended as a walk-up-and-use system, participants were provided with a grounding introductory spiel and an experimenter-directed training task. After the completion of those tasks, the participants were asked to think-aloud while completing the remaining experimental tasks. The experimenter provided no assistance to the participants during the completion of the tasks.

We recruited N=12 administrator volunteers for an experimental session lasting about 90 minutes. They were paid \$15/hour for their time. Volunteers were asked if they had experience with programming, and those who did were disqualified from participation. We began by introducing the tool and acquainting the users with its goals and concepts. The experimenter then introduced participants to the concept of the thinkaloud experimental setup.

Next users completed a preliminary survey detailing their background level of computer usage and expertise. To ensure that users understood the task we asked them to take three minutes to complete as much of a task as they could manually, specifically by copying and pasting directly from the Association for Computing Machinery (ACM) website into a spreadsheet.

Then, to illustrate the use of the system, the participant performed a representative Mixer task with minute direction from the experimenter. The training task was:

- Task 0: Find all researchers from Institution X who published in the latest conference of Conference Z

Then, one by one, we asked them to respond to messages in a pre-loaded email account. Each message contained a request from a contrived boss for the completion of an experimental task; users indicated completion of the task by replying to the email with their best attempt at the answer. During the completion of the tasks, the participants' actions and audio were recorded using Camtasia for later analysis.

- Task 1: Find all researchers from Institution Y who published in the latest conference of Conference W
- Task 2: Find all coauthors of Researcher R in the last three years
- Task 3: Find email addresses for all members of Club C
- Task 4: Find all coauthors of Researcher S in the last three years

The tasks were mostly from the ACM domain in order to minimize the amount of domain knowledge presupposed or learned in-experiment on the part of the participant. Task 1 was chosen as an isomorph of the demonstration task to cement the participant's understanding of the process of extracting a subset, then using spreadsheet functionality to select the appropriate subset. Task 2 has the same form, but the web interactions are novel, sometimes changing which pieces of information require a new server response. Task 4 is a repeat of Task 2 with different parameters, but introduces a minor complicated factor that Researcher S's first paper is published alone (i.e. the only coauthor is the first author). Task 3 is completely novel in the sense that the output of one website is used as the input of another. Participants were not instructed how to use Mixer to combine data from multiple websites, nor were they alerted that Task 3 had any characteristic different from the rest.

Additionally, Tasks 2 and 4 had two different solution paths. The ACM listing of an author's publications lists all publications with links to pages about the individual papers; alongside the link is a listing of metadata about the paper including authors and publication date. Thus the problem may be solved within a single page, since all needed information is present in the page. Alternately, the problem can be solved by extracting the required metadata from each publication's page in turn.

Following completion of the tasks, participants answered a post-study questionnaire containing the TAM3 [37] (Technology Acceptance Model 3) instrument. TAM3 measures a new technology's perceived usefulness and perceived ease of use. Previous research shows a strong relationship between these two perceptions and eventual system use. TAM3 responses were made on a 7-point Likert scale (1 = "extremely unlikely to use," 4 = "neither," 7 = "extremely likely to use").

4 Findings

After the participant had correctly communicated the desired behavior to Mixer, Mixer turned control back over with a filled table of data. In 42 (about 88%) of the tasks the participant was able to correctly filter the data and direct the completed form to the experiment's simulated boss. In one case, a participant was unable to do so for Task 1 and gave up; that same participant was able to correctly marshal the data in the subsequent tasks. In four cases, participants needed one or two more attempts to effect the correct answer. In one case, a participant needed five attempts.

Our 12 participants successfully completed all tasks. They eventually constructed all of the necessary tables with the agent's help. Because the experiment required the participants to think aloud as they worked, the amount of time participants took to accomplish the task is not meaningful; instead, we record the number of attempts they made before they were able to productively turn control over to Mixer. We counted an attempt every time the participant started over with a fresh workspace during the completion of a task. All participants' number of attempts are presented in Figure 2.

Participants exhibited some difficulty constructing a workspace table containing all of the information required to complete the task. 17 task attempts failed due to missing query attributes, for example by failing to include students' names when only email addresses were strictly required. Two participants successfully extracted a spreadsheet, only to find that missing selection attributes precluded them from sorting and filtering down to the correct answer. They immediately re-demonstrated the correct workspace without error.

Two participants constructed a table without an attribute explicitly requested, then corrected their oversight. Of the overlooked attributes of all categories, only one instance occurred in the final task. Several participants appeared to struggle with Mixer's expectation that the user would only provide information in the first row, leaving Mixer in charge of filling subsequent rows. Three participants, all in Task 3, tried to continue filling subsequent rows before clicking on "Fill Table."

Two participants, again in Task 3, attempted to search multiple email addresses at once by pasting all the names, separated by spaces, into the search box; in response, the directory application returned no results and the participants started over. One participant attempted to explicitly select a column of attributes from the target page.

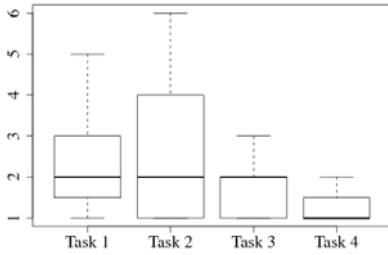


Fig. 2. Number of attempts to construct the correct workspace

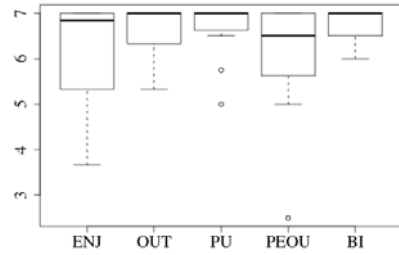


Fig. 3. Participants' ratings of Mixer along TAM constructs

Another common breakdown occurred with respect to participants' decision to invoke the program by pressing the "Fill Table" button. One participant chose to restart Task 1 after exporting a table with unfilled cells, i.e. failing to invoke the program at all. Four participants discovered a solution to Task 2 that did not require the use of the "Fill Table" button to complete the task. They then expressed confusion that the "Fill Table" button was inactive. Two of the confused participants precipitously restarted after encountering the confusion. All four participants used the same approach on Task 4 and did not hesitate to complete the task without using the "Fill Table" button. Three additional users discovered this approach while completing Task 4. Each expressed confusion that the invocation button had no effect, but all pressed onward and successfully completed the task.

Eight participants encountered the dialog box warning against typing in a query field. One successfully circumvented the dialog, typing in the information and thereby causing the task attempt to fail. One user attempted to select data not recognized by the wrapper and send it to Mixer.

Figure 3 shows participants' responses to the TAM instrument. The Cronbach alpha scores, all above 0.8, indicate that the scores are internally reliable, in the sense that answers to multiple questions seem to measure the same underlying construct.

Many users expressed pleasant surprise at the capabilities of Mixer, using adjectives like "cool", "awesome", and "brilliant." One user said "I want this program. Even if it can't find everybody" and another asked "When can I get this?" All of these laudatory quotes came immediately after the user was able to successfully complete Task 3. Two users praised the visibility of Mixer's practice of showing each visited page as the table is filled. Two participants expressed displeasure at the use of Google Docs for the spreadsheet export; they claimed to be more proficient with Microsoft Excel. During the closing interview, several participants inquired when Mixer would be available to them for their jobs.

5 Discussion

Administrators could successfully use Mixer to automate tedious information retrieval tasks. They were successful at creating the first row of a table as a way of communicating to an agent the information they wanted. At first, many administrators

struggled to complete a task. Sometimes this was caused by software bugs in Mixer, and sometimes it was caused by participants struggling to conceive of tables the agent could act upon. The agent often needs more context (additional columns) to complete an action than the administrator strictly needs to complete their task. Administrators struggled with including this context, indicating that, initially, they had trouble seeing the problem from the agent's perspective. However, the reduction in the number of attempts needed to successfully complete a task from the first task to the fourth task (Figure 2) provides some evidence that administrators can quickly learn to conceive of the tables in a way that allows the agent to assist them with their task.

Mixer supports two types of tasks: repeated retrieval from within a single data source, and retrieval from across more than one data source. We expected that working with more than one data source would be more difficult, but we did not see evidence for this. A comparison of the number of attempts made on Task 3, which required participants to connect two data sources, to the other three tasks that all use a single data source (Figure 2) does not indicate that participants found the multiple data-source task more difficult. In addition, we see nothing in the utterances during the thinkaloud to indicate that participants made any kind of distinction between these tasks. Though administrators struggled with the fact that Mixer requires copying and pasting into forms instead of typing, they did not seem to find demonstrating a link between data sources challenging. This finding is especially interesting in the light that, as indicated above, participants were not supplied with any hints to how to effect a link between multiple data sources.

The use of wrappers to augment the target page with a highlight indicating a legally selectable element provides a design advance. The highlights were intended to help users understand the limitations of the agent's communication ability and to help negotiate the problem space between human and agent. In our previous iterations of the interface, we allowed users to copy and paste from the target page directly into the workspace table, and this often lead to breakdowns. Participants seemed to have no trouble understanding how to use these highlights and never expressed any utterances or opinions that this limited their ability to use Mixer to automate their work. We think this technique could be used in many other mixed-initiative interfaces, where users struggle to understand the scope of what the agent can do.

Mixer's interaction design specifically avoids the challenge of precise specification: the need to communicate that the user wants only a subset from within a larger set of data. Instead, the agent retrieves the larger set and encourages users to use a spreadsheet to filter down to the precise information. The fact that participants were able to correctly do so using a spreadsheet corroborates the notion that end users can conceive of the task as a nested table, and furthermore that the unnesting of the table into a spreadsheet table is an intuitive concept for administrators.

TAM produced very high ratings for Mixer for Perceived Ease of Use, Perceived Usefulness, and Behavioral Intention (to use). Scores of 6 or above on a seven-point scale give us confidence that administrators recognize the value of automating their tedious information retrieval tasks and that they would likely use Mixer in their work. Additionally, users gave a high TAM score to the quality of Mixer's output. This may be due in part to Mixer showing exactly which pages contributed to the result as it filled in the table. Several participants singled out this aspect independently for praise.

Mixer's interaction design specifically deemphasizes and to a certain extent even hides the fact that users are engaged in a programming task when they construct the first row of the table. This is a radical departure from most work in the web PBD community. We speculate that systems that similarly deemphasize the programming aspect of the task will generally be more likely to succeed with nonprogrammers. Much more work would need to be done to rigorously evaluate this speculation, though the fact that administrators with no programming experience could successfully use Mixer and their high TAM scores reflect positively, as does the fact that nonprogrammers have struggled with other PBD systems. In the same vein, more work would need to be done to show that the administrators' resistance to programming stems from the sense that the work is perceived to fall outside the scope of what their job should be. We can suggest that this is a rich direction for further research on PBD interfaces.

In terms of PBD, Mixer embraces the notion that administrators would be disinclined to use a tool that feels like programming. The most obvious Mixer design decision in this line is that the user does not see the program as lines of code, nor as an equivalent data-flow representation of the procedure. A more subtle example is the way Mixer enforces copying and pasting as, from the user's perspective, an arbitrary constraint, rather than asking the user to understand the programming concept of using a variable as opposed to its value. Consequently, in terms of PBD systems, Mixer is near one extreme of the spectrum, ranging from those that expect the user to construct an explicit model of the program as a sequence of low-level actions, to those that do not. The success of our participants in using Mixer, as well as their recognition of its relevance and applicability to their jobs, seems to lend credence to this notion: the less end users feel like they are engaging in "programming" while using a PBD system, the less likely they seem to be to eschew the system as unrelated to their realm of responsibility.

6 Related Work

Nardi [27] notes the widespread use of sophisticated forms in human-human interaction, and the surprising facility of nontechnical people in rapidly learning and making use of them. She also observes that users can more readily assimilate new formal representations when they have a preexisting interest or job-related requirement to do so. Rode et al [31] note the same phenomenon. They note the similarity between ovens and VCRs in terms of programming. Abstractly, both devices allow users to instruct a device to turn on at a specified time and run for a specified duration, and to set the state of a specific feature: the channel of the VCR and the temperature of the oven. Surprisingly, despite the indistinguishability of the tasks at the abstract level, they found a pronounced gender difference in users' abilities to perform the tasks. Women, who generally exert more control over the kitchen, had more success programming ovens, and conversely, men, who generally exert more control over entertainment devices in the home, had more success programming VCRs. These research results lead us to speculate that one reason office workers have not readily accepted PBD systems is that they cast the task as "programming": a type of work that generally falls outside the common social role

description for an administrator. Mixer addresses this by specifically disguising the fact that the administrator is programming when interacting with the tool.

Malone et al note the potential of semi-structured forms as a means of expressing human practice and intention in a manner that is amenable to agent assistance [23]. Their work focuses on structuring email conversations so that agents can assist in the coordination of human activities, essentially providing a mechanism for the agent to eavesdrop on the human communication. VIO [40] complements Malone by providing the reverse: a form mechanism whereby users are given insight into the actions of the agent, and hence the opportunity to identify and repair agent errors.

Nardi and Miller [28] build on the work of Lewis and Olson [21] in singling out spreadsheets, which can be viewed as frameworks for the creation of ad hoc forms, as an emblematic context where people routinely "program", in the sense that they induce nontrivial computational behavior. Nardi and Miller delineate several specific aspects of spreadsheets which render them particularly acceptable to end user interviewees. First, the computational paradigm of spreadsheets matches the way the end user conceptualizes the task; Norman [29] characterizes this alignment as bridging the "Gulf of Execution" between the user's conceptualization of the goal and the system's formalism. In particular, the high-level functions provided by the spreadsheet shield the user from the difficult task of "synthesizing" the desired functionality from simpler primitives. Secondly, spreadsheets compactly represent the entire task in a single tabular view, often on a single screen.

Our previous Wizard-of-Oz study of Mixer demonstrated that these advantages of spreadsheets apply to administrators approaching data integration tasks, specifically pointing out the conceptual alignment between user and agent as well as the unified nature of the shared table representation. Several other systems settle on a similar tabular interface between the user and an observing PBD web data integration agent. Vegemite [22] asks the user to create a set of "VegeTables," each of which corresponds to a script for combining two websites. Karma [36], Dontcheva et al [11] and Mashroom [38] build separate tables for each extracted website; additionally, Mashroom explicitly uses nested tables (specifically with an eye towards comprehensibility by end users). Each of these systems asks the user to explicitly "merge" extractions from different websites into a coherent table. In contrast, Mixer encourages the user to construct the single, unified table that seems to match her underlying conceptualization of the task. This spares the user the confusion inherent in synthesizing, or merging, the results of the various subtasks together. As a consequence, Mixer enables users to construct integration tasks over one website, or over several websites, without necessarily observing the distinction.

Mixed-initiative research focuses on advancing methods for collaboration between computer agents and people where each party has its own knowledge, ways of reasoning, and abilities to understand and act in order to advance toward a common goal [1, 14]. Many issues remain to be answered, including several interrelated needs with respect to interaction between agent assistants and people [33]:

- Awareness: knowledge of problem and goal must be shared by human and agent
- Task: roles and responsibilities must be shared between human and agent
- Communication: both human and agent must be able to express knowledge and needs.

PBD interfaces present a particular challenge with respect to the awareness issue: the user and the system have a fundamental mismatch with respect to the goal of the interaction. The central goal of a PBD system is to infer a program from the user's actions; for the user the construction of the program is subsidiary, at best, to the goal of completing some task. As noted above, Rode et al [31] observe that users are far less successful in performing programming tasks outside their perceived area of responsibility. Consequently, Mixer explicitly attempts to avoid presenting the user with tasks that feel like programming.

The task issue concerns the division of action between humans and agents. The principal actions of a PBD session [18] are program demonstration (or creation), program invocation, and program execution. Mixer incrementally constructs a program by observing all actions taken within the browser, from the time that the user invokes Mixer to the time that the user presses a button to invoke the demonstrated program. Mixer then executes the program. Thus Mixer presents a strong distinction between user actions (before invocation) and system actions (after invocation). This separation of activity is stricter in Mixer than in some PBD systems, such as Eager [8], which assist the user in deciding when to invoke the observed behavior.

The communication issue arises in a couple of ways from what Cypher [9] calls the classic challenges of PBD: (1) inferring the user's intent; and (2) presenting the created program to the user. The first challenge concerns the user communicating with the system via the demonstrated actions, and the second challenge concerns the system communicating the recorded action sequence to the user.

The first challenge arises because the user's actions usually insufficiently delineate a unique program, a point illustrated by Lau et al with an explicit version space argument [19]. PLOW [2] receives richer input from the user by eliciting and utilizing natural language explanations for the user's actions. Wrangler [16] asks the user to select after each action the statement in the implementation language corresponding to the level of generalization required. Rather than eliciting additional input from the user, Mixer overcomes the problem by exploiting rather strong simplifying assumptions about the types of problems Mixer is expected to solve.

The user has the responsibility to demonstrate their knowledge of a single row of the table, and Mixer assumes full responsibility for inferring the best possible procedure from that demonstration. Although the user need not understand the workings (or even the existence) of the program, the user does need to be aware that the agent is observing; in other words, the user is expected to take an "intentional stance" [10, 24] with respect to showing Mixer how to perform the desired task. Mixer asks the user to intentionally demonstrate similar information to that detected automatically by TX2 [4].

As to the second challenge, Modugno and Myers [25] further delineate the communication role played by the program in PBD systems, as a list of opportunities presented to the user:

1. the user can confirm that the program will behave as desired;
2. the user can correct or generalize the program; and
3. the user can store all or part of the program for later use or modification.

Mixer provides limited information about the inferred program through the intermediate depiction of the workspace, giving the user implicit confirmation responsibility as well as some ability to correct unexpected columns in the workspace.

Although many PBD systems outside the web context communicate the program in forms other than as lines of code, the code approach is the most common in web PBD systems. Chickenfoot [5] records web actions as general JavaScript. CoScripter [20] chooses a slightly more user-friendly approach, representing the program in a “sloppy” or natural programming language. Query-by-Example and Office-by-Example [41] utilize a form as a shared communication structure, but require a user to understand and specify programmatic variable structure within the forms. Mixer uses a single nested table form as the principal communication medium between the human and the agent, which diminishes the variety of programs Mixer can produce but dramatically simplifies the user’s interaction with the system.

Over the last few years there has been a great amount of research interest in streamlining the process of creating web mashups [3]. By focusing on ad hoc reports rather than mashups (i.e. the output rather than the program), Mixer differs philosophically from many mashup projects; in particular, Mixer aims to allow users to conceptualize data integration problems uniformly, whether or not some pieces of information lie across web server boundaries. Whereas mashup systems emphasize reusability and generality, Mixer focuses on how administrators can retrieve and integrate the types of data they need for their jobs.

Nevertheless, Mixer shares some overlap with mashup systems in that Mixer presents a user-friendly solution to the source modeling and data integration problems, with particular attention to the database *joins*. Thus, Mixer could coexist in a mashup ecosystem with user-appropriate solutions to wrapper generation (e.g. reform [35] or the summaries of Dontcheva et al [12]) or data cleaning (e.g. Potters Wheel [30] or Potluck [15]). Mash Maker [13] provides a representative mashup ecosystem, distinguishing between end user-specified wrappers and developer-provided widgets, which combine and visualize the wrapped data. In this perspective, Mixer presents a mechanism for nonprogrammers to create useful widgets without developer intervention.

7 Conclusion

Mixer advances mixed-initiative PBD interaction through a novel user-constructed nested table communication method that allows users to declare the outcome they want while implicitly demonstrating how the agent should programmatically perform the task. Mixer specifically allows administrators to automate repetitive web data retrieval and integration tasks they find to be tedious to perform. Our evaluation of the system shows specifying the table to be an effective method for people and the agent to communicate their varying knowledge and needs. The evaluation also reveals a strong likelihood that administrators would use Mixer if it were available to them. The interaction presented in Mixer represents a transition in how office workers engage in computing. Instead of forcing workers to rely on their ability to adapt to the design of IT systems, Mixer empowers workers to leverage their expertise in web data retrieval to train agents to undertake tedious information integration tasks for them.

This work is supported by grant number H133E080019 from the United States Department of Education through the National Institute on Disability and Rehabilitation Research.

References

1. Mixed-Initiative Interaction. *IEEE Intelligent Systems* 14, 14–23 (1999)
2. Allen, J., Chambers, N., Ferguson, G., Galescu, L., Jung, H., Taysom, W.: PLOW: A Collaborative Task Learning Agent. In: *AAAI*, pp. 22–26 (2007)
3. Beemer, B., Gregg, D.: Mashups: A Literature Review and Classification Framework. *Future Internet* 1(1), 59–87 (2009)
4. Bigham, J.P., Kaminsky, R.S., Nichols, J.: Mining web interactions to automatically create mash-ups. In: *UIST*, pp. 203–212 (2009)
5. Bolin, M.: *End-User Programming for the Web*, Masters Thesis. MIT (2005)
6. Boodman, A.: Greasemonkey, <http://www.greasespot.net/>
7. Clark, J., DeRose, S.: XML Path Language (XPath) Version 1.0, W3C (1999)
8. Cypher, A.: Eager: Programming Repetitive Tasks by Demonstration. In: *Watch What I Do: Programming by Demonstration*, pp. 205–217. MIT Press, Cambridge (1993)
9. Cypher, A.: End User Programming on the Web. In: *No Code Required: Giving Users Tools to Transform the Web*, pp. 3–22. Morgan Kaufmann, San Francisco (2010)
10. Dennett, D.C.: *The Intentional Stance*. The MIT Press, Cambridge (1987)
11. Dontcheva, M., Drucker, S.M., Salesin, D., Cohen, M.F.: Relations, cards, and search templates: user-guided web data integration and layout. In: *UIST*, pp. 61–70 (2007)
12. Dontcheva, M., Drucker, S.M., Wade, G., Salesin, D., Cohen, M.F.: Summarizing personal web browsing sessions. In: *UIST*, pp. 115–124 (2006)
13. Ennals, R., Brewer, E., Garofalakis, M., Shadle, M., Gandhi, P.: Intel Mash Maker: join the web. *SIGMOD Rec.* 36(4), 27–33 (2007)
14. Horvitz, E.: Reflections on Challenges and Promises of Mixed-Initiative Interaction. *AI Magazine* 28(2) (2007)
15. Huynh, D., Miller, R., Karger, D.: Potluck: Data Mash-Up Tool for Casual Users. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) *The Semantic Web. LNCS*, vol. 4825, pp. 903–910. Springer, Heidelberg (2007)
16. Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler: Interactive Visual Specification of Data Transformation Scripts. In: *CHI* (2011)
17. Kawanaka, S., Borodin, Y., Bigham, J. P., Lunn, D., Takagi, H., Asakawa, C.: Accessibility commons: a metadata infrastructure for web accessibility. In: *Assets*, pp. 153–160 (2008)
18. Kosbie, D.S., Myers, B.A.: PBD Invocation Techniques: A Review and Proposal. In: *Watch What I Do: Programming by Demonstration*, pp. 415–422. MIT Press, Cambridge (1993)
19. Lau, T., Wolfman, S.A., Domingos, P., Weld, D.S.: Programming by Demonstration Using Version Space Algebra. *Mach. Learn.* 53(1-2), 111–156 (2003)
20. Leshed, G., Haber, E.M., Matthews, T., Lau, T.A.: CoScripter: automating & sharing how-to knowledge in the enterprise. In: *CHI*, pp. 1719–1728 (2008)
21. Lewis, C., Olson, G.: Can principles of cognition lower the barriers to programming? In: *Empirical Studies of Programmers: Second Workshop*, pp. 248–263. Ablex, Norwood (1987)

22. Lin, J., Wong, J., Nichols, J., Cypher, A., Lau, T.A.: End-user programming of mashups with vegemite. In: *IUI*, pp. 97–106 (2009)
23. Malone, T.W., Grant, K.R., Lai, K.-Y., Rao, R., Rosenblitt, D.: Semistructured Messages Are Surprisingly Useful for Computer-Supported Coordination. *ACM Trans. Inf. Syst.* 5(2), 115–131 (1987)
24. Mausby, D., Witten, I.H.: Metamouse: An Instructible Agent for PBD. In: *Watch What I Do: Programming by Demonstration*, pp. 155–181. MIT Press, Cambridge (1993)
25. Modugno, F., Myers, B.: Graphical Representation and Feedback in a PBD System. In: *Watch What I Do: Programming by Demonstration*, pp. 415–422. MIT Press, Cambridge (1993)
26. Myers, B.A.: Peridot: Creating User Interfaces by Demonstration. In: *Watch What I Do: Programming by Demonstration*, pp. 125–154. MIT Press, Cambridge (1993)
27. Nardi, B.A.: A small matter of programming: perspectives on end user computing. MIT Press, Cambridge (1993)
28. Nardi, B.A., Miller, J.R.: *The Spreadsheet Interface: A Basis for End User Programming*. HP Laboratories (1990)
29. Norman, D.A.: *Cognitive Engineering*. In: *User Centered System Design: New Perspectives on Human-Computer Interaction*, pp. 31–61. Lawrence Erlbaum Associates, Mahwah (1986)
30. Raman, V., Hellerstein, J.M.: Potter’s Wheel: An Interactive Data Cleaning System. *The VLDB Journal*, 381–390 (2001)
31. Rode, J.A., Toye, E.F., Blackwell, A.F.: The fuzzy felt ethnography—understanding the programming patterns of domestic appliances. *Personal and Ubiquitous Computing* 8(3), 161–176 (2004)
32. Sundström, P., Taylor, A.S., Grufberg, K., Wirström, N., Belenguer, J.S., Lundén, M.: Inspirational Bits: Towards a shared understanding of the digital material. In: *CHI* (2011)
33. Tecuci, G., Boicu, M., Cox, M.: Seven Aspects of Mixed-Initiative Reasoning: An Introduction to this Special Issue on Mixed-Initiative Assistants. *AI Magazine* 28(2) (2007)
34. Tomasic, A., Zimmerman, J., Hargraves, I., McMullen, R.: User Constructed Data Integration via Mixed-Initiative Design. In: *Interaction Challenges for Intelligent Assistants*, pp. 122–123 (2007)
35. Toomim, M., Drucker, S.M., Dontcheva, M., Rahimi, A., Thomson, B., Landay, J.A.: Attaching UI enhancements to websites with end users. In: *CHI*, pp. 1859–1868 (2009)
36. Tuchinda, R., Szekely, P., Knoblock, C.A.: Building Mashups by example. In: *IUI*, pp. 139–148 (2008)
37. Venkatesh, V., Bala, H.: Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences* 39(2), 273–315 (2008)
38. Wang, G., Yang, S., Han, Y.: Mashroom: end-user mashup programming using nested tables. In: *WWW*, pp. 861–870 (2009)
39. Zimmerman, J., Rivard, K., Hargraves, I., Tomasic, A., Mohnkern, K.: User-created Forms as an Effective Method of Human-agent Communication. In: *CHI 2009*, pp. 1869–1878 (2009)
40. Zimmerman, J., Tomasic, A., Simmons, I., Hargraves, I., Mohnkern, K., Cornwell, J., McGuire, R.M.: VIO: a mixed-initiative approach to learning and automating procedural update tasks. In: *CHI*, pp. 1445–1454 (2007)
41. Zloof, M.M.: QBE/OBE: A Language for Office and Business Automation. *IEEE Computer* 14(5), 13–22 (1981)

Speaking to See: A Feasibility Study of Voice-Assisted Visual Search

Victor Kaptelinin^{1,2} and Herje Wåhlen²

¹ University of Bergen, Department of Information Science and Media Studies,
P.O. Box 7802, 5020 Bergen, Norway

² Umeå University, Department of Informatics, 901 87 Umeå, Sweden
vka062@uib.no, herjew@gmail.com

Abstract. The paper presents the concept, implementation, and a feasibility study of a user interface technique, named VAVS (“voice-assisted visual search”). VAVS employs user’s voice input for assisting the user in searching for objects of interest in complex displays. User voice input is compared with attributes of visually presented objects and, if there is a match, the matching object is highlighted to help the user visually locate the object. The paper discusses differences between, on the one hand, VAVS and, on the other hand, voice commands and multimodal input techniques. An interactive prototype implementing the VAVS concept and employing a standard voice recognition program is described. The paper reports an empirical study, in which an object location task was carried out with and without VAVS. It was found that the VAVS condition was associated with higher performance and use satisfaction. The paper concludes with a discussion of directions for future work.

Keywords: Voice recognition, visual search, multimodal input, voice command.

1 Introduction

Visual search is a crucial component of a wide range of interactions between people and digital technologies; it involves scanning displayed information to detect the presence of an object of interest, identify its location, or explore object’s properties. For instance, if the user wants to make sure that the last email message from a certain customer has been actually answered, the user may scan the list of messages in the Inbox window and search for the last message from the client, visually locate the message line, and check whether the icon on the left contains a small arrow. Finding a certain street on a digital map, looking up information about a flight on a “Departures” monitor, and many other everyday interactions with electronic displays are critically dependent on visual search. Visual search may or may not involve carrying out an action with the object of interest.

In this paper we argue that for users of digital technologies visual search may be associated with certain problems, and that there is a need to provide the users with more advanced technological support for visual search. We introduce a user interface technique, named VAVS (voice-assisted visual search), which aims to facilitate visual search by employing user’s voice input for visually highlighting objects of interest.

In the remainder of this paper we present the rationale behind the VAVS technique, discuss how the technique is related to previous work, describe an interactive prototype of a system implementing the technique, and report a feasibility study, in which the prototype was employed in an object location task.

2 Background

Making a large number of information objects simultaneously available to the user for viewing has important advantages. In particular, it decreases the need for the user to open and “look inside” opaque containers, such as folders or pull-down menus to find objects of interest [2]. However, these advantages come with a price. In case of dense, complex displays, when the object of interest (the “target”) is presented simultaneously with a large number of other objects (“distractors”), visual search becomes a more demanding task [9]. Problems with visual search can be aggravated by several factors, such as users’ age (children and the elderly have more difficulties than young adults), level of stress, and certain health conditions, as well as how specifically the target is defined when a person carries out a visual search task (e.g., [4, 9]). The problems are likely to worsen in the future, since the screen size and resolution of computer monitors, public information displays, tabletops, and so forth, are ever-increasing, which means displaying more (and more complex) information objects.

Helping users visually identify their objects of interest has always been high on the agenda of the design of graphical user interfaces. Well-designed interfaces visually emphasize potentially important objects and de-emphasize less important ones [2, 9]. Relative visual salience of displayed objects can be a static feature of an interface or it can dynamically change depending on the task context (for instance, the default button in a dialogue window is highlighted to make it easier for the user to choose the most likely option).

These strategies for supporting users’ visual search seem to have been often successful in the past and they remain to be useful. However, they are, arguably, not sufficient for addressing current challenges. Making potentially relevant objects visually salient does not scale up to complex displays and complex tasks. If all potentially relevant objects are visually salient, their absolute number can be overwhelming. In addition, when a large amount of information is displayed, it might be difficult for the system to anticipate just what objects can be of importance to a particular user in a particular context and, therefore, should be visually emphasized.

These logical arguments are consistent with the evidence obtained in empirical studies. For instance, Andrews et al. [1] describe “losing the cursor” and users’ confusion caused by “windows and dialog boxes opening or gaining focus in unexpected locations” as common problems with large displays.

To address the problems, discussed above, we have developed a user interface technique for assisting the user in searching for objects of interest in complex displays. The underlying idea of the technique, named VAVS (voice-assisted visual search), is employing user’s voice input for guiding user’s visual attention.

Figure 1 shows an overall structure of a VAVS-enabled interface. The user scans an image displayed on a screen (S) to locate a certain object. The user can also use a

microphone (M) to describe object's attributes, such as its name. The voice input is processed and compared with attributes of objects displayed on the screen and, if there is a match, the matching object is visually highlighted. For instance, if a person, looking at a map of Colorado on a computer display, is saying "Hmm... Mancos... Mancos...", the location of the town on the map is temporarily highlighted.

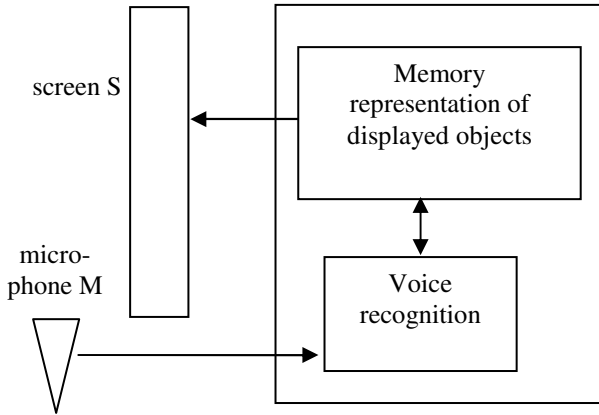


Fig. 1. Overall structure of a VAVS-enabled interface

The VAVS technique should be differentiated from two other ways of using user's voice input, which have been actively explored in previous research: voice commands (in a broad sense, including voice-based queries) and multi-modal input techniques.

Like voice commands, which are an increasingly common interaction technique, for instance, in in-car systems [5], VAVS also employs users' voice input. Unlike voice commands, however, VAVS does not cause substantial changes in the state of the system. Its effect is limited to visually highlighting potential objects of interest. If a user's voice input results in highlighting some other object than the desired one (either because of a user's mistake or system's misinterpretation) the user can simply ignore the highlighting when proceeding with their task. It also means that VAVS users do not have to be overly concerned about negative consequences of their mistakes (while users of voice command systems have to overcome a substantial initial barrier before they start to feel comfortable with a system [5]).

A related approach to employing user's voice in human-computer interaction is supporting multi-modal input, or multi-modal dialogue, that is, enabling the use of voice input in combination with other interaction modalities [3,7,8]. An example of this approach is the classic "Put-that-there" system [3], which combines voice and gesture. For instance, to move an object across a large display the user specifies a command by voice (i.e. by saying "put"), points to an object and selects it by saying "that", and finally indicates a new location by pointing to it and saying "there".

Users of the "Put-that-there" system, as well as users of more recent systems that implement the same general approach [8], need to know—in advance—the spatial locations of objects of interest and convey these spatial locations to the system when instructing it to carry out a desired action. Support for selecting an object to indicate

the system what it should act upon (cf. Windows Speech Recognition [10]) may partially overlap with support of visual search, but the general approach adopted by VAVS is, in a sense, opposite. According to that approach, it is the system that conveys the spatial locations of objects of interest to the user, rather than the other way around. Accordingly, a VAVS system has a number of features differentiating it from multimodal input systems. For instance, all potential objects of interest, rather than just potential objects of actions, should be “highlightable”.

The next section presents a feasibility study intended to gain empirical evidence on whether VAVS can be helpful when supporting users in finding objects of interest on complex displays.

3 Method

Participants. Eight university students, native Swedish speakers and fluent English speakers, 23 to 33 years old, took part in the study.

Procedure. The participants were tested individually. Each session started with a profile calibration procedure that took five to twelve minutes. After that each participant was presented with a series of object location tasks. In each task a participant was presented with a name of a map region in the top left corner of the screen and was required to locate and click the corresponding map region using the mouse. The user had to click the correct map region to proceed to the next task. Each participant was presented with 96 object location tasks divided into two blocks. One of the blocks corresponded to the “VAVS” condition (voice input was enabled), and the other block corresponded to the “non-VAVS” condition (voice input was disabled). In each block the first five tasks were practice tasks, not included in the analysis. Finally, the participants were briefly interviewed about their experience with VAVS. The duration of a typical session with a participant was about 30 min.

Equipment. The hardware used in the study was an Apple MacBook Pro computer (15-inch, 2.33 GHz Intel Core 2 Duo processor, 4 GB SDRAM) running Mac OS X 10.6.3, connected to two external devices: Microsoft IntelliMouse Explorer 3.0 and Logitech USB Desktop Microphone.

Prototype. An interactive prototype of a VAVS-enabled system was developed for the study in AppleScript and JavaScript. The prototype was integrated with a speech recognition program, Nuance MacSpeech Dictate International, version 1.5.8. The visual interface was implemented as an HTML document opened in full screen mode.

The functionality of the prototype included: (a) displaying a map featuring a number of regions (“countries” or “states”), (b) displaying the name of one of the map regions in the top left corner of the screen, (c) measuring the time interval between presenting a name of a region and a mouse click on the corresponding map region, (d) recognizing a map region name uttered by the user, and (e) visually highlighting the map region corresponding to the name. In the control (“non-VAVS”) condition functions (d) and (e) above were disabled.

Materials. Two maps, loosely based on Adobe Photoshop filter-generated images as reference for map region borders, were created for the study. *Map A* was derived from

a map of Europe, and real English names of European countries were randomly assigned to different map regions (see Figure 2). *Map B* was derived, in a similar manner, from a US map.

The maps were designed to make sure the participants were familiar with the names of the map regions but could not use their previous knowledge to infer the locations of map regions from their names. Therefore, the participants had to visually scan the maps in order to complete the experimental tasks.



Fig. 2. An adapted fragment of Map A (“Liechtenstein” is visually highlighted)

Design. The study employed a one-factor within-subject design, with the independent variable being Voice Input (“VAVS” condition vs. “non-VAVS” condition). The main dependent variable was task completion time.

The design was balanced to minimize the potential effects of condition sequence and map types. The participants were divided into two equal sub-groups. The first sub-group completed the first block of tasks in the “VAVS” condition and the second block in the “non-VAVS” condition; for the second sub-group the sequence was the opposite. In each of these two sub-groups half of the participants worked with Map A in the “VAVS” condition and Map B in the “non-VAVS” condition, while for the other half the correspondence between maps and conditions was the opposite.

4 Results

As mentioned, the experiment procedure required a task to be correctly completed before the next task could be presented. All participants were able to complete all tasks in both conditions, which allowed us to use time to correctly complete a task as

an integral performance indicator, in which error costs, both participants' mistake and voice recognition errors, were reflected as added "error time".

Voice recognition error rate in the VAVS condition—calculated as the percentage of tasks, in which the participants had to pronounce a state or country name more than once—was 19%. In two cases the experimenter had to intervene and suggest the right pronunciation (while the tasks were performed by the participants themselves). A likely reason for the high error rate was that native Swedish speakers were asked to pronounce English words.

Figure 3 shows accumulated times for completing blocks of tasks in the two experimental conditions for each of the eight participants. Figure 3a shows the results of the four participants (S1, S2, S3, and S4), who worked with Map A in the "VAVS" condition and Map B in the "non-VAVS" condition. Figure 3b shows the results of the four participants (S5, S6, S7, and S8), who worked with Map B in the "VAVS" condition and Map A in the "non-VAVS" condition.

The results, shown in Figure 3, indicate that in the "VAVS" condition *each* of the participants completed the experimental tasks faster than in the "non-VAVS" condition. While the average accumulated task completion time in the "non-VAVS" condition was **384** seconds; in the "VAVS" condition it was **176** seconds.

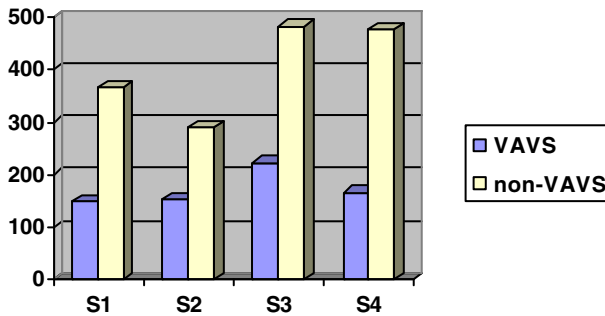


Fig. 3a. Accumulated task completion times, in seconds, for the experimental conditions of the study. Participants: S1, S2, S3, and S4. "VAVS": Map A, "non-VAVS": Map B.

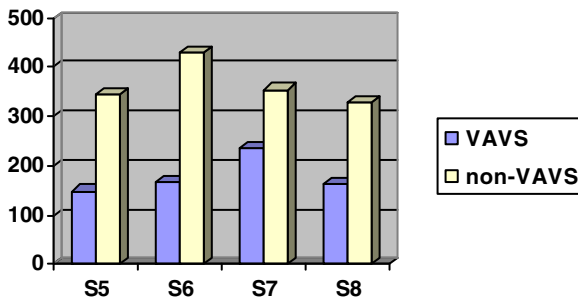


Fig. 3b. Accumulated task completion times, in seconds, for the experimental conditions of the study. Participants: S5, S6, S7, and S8. "VAVS": Map B, "non-VAVS": Map A.

The results were analyzed using the Wilcoxon signed-rank test. The difference between the “VAVS” and “non-VAVS” condition was found to be statistically significant ($N=8$, $W_+=36$, $W_-=0$, $p=.005$).

In their interview comments all participants indicated that they were positive about the VAVS technique and wanted it to be used in a diversity of everyday contexts.

5 Discussion of Results and Future Work Directions

The results of our study suggest that employing user voice input for visually highlighting objects of interest can be associated with higher performance and positive user experience. Given that the use of voice at the user interface is complicated by a number of factors [2], and speech-based interfaces have been, in general, much less successful than it was anticipated in the past [6,7], we consider the results of our study encouraging. The study also showed that a standard speech recognition program can be accurate and reliable enough to support VAVS-enabled interaction.

It should be noted that advantages of VAVS were observed in conditions in which the advantages were not self-evident. The participants had to speak a foreign language, which was probably one of the reasons behind the high voice recognition error rate and, consequently, resulted in higher task completion times in the VAVS condition. In addition, the map image used in the study was relatively simple, which meant that unassisted visual search remained a viable option. It is reasonable to assume that if the users spoke their native language and worked with large displays and complex images, VAVS’ advantages would be even more significant.

Can the findings be explained by a “negative familiarity” effect, that is, by target familiarity being an impediment rather than help in the specific task used in the study? If this explanation is correct, the findings from our study are only valid for rare instances of search tasks. However, the results do not support this hypothesis: if it were correct, the longest search time would be for “Sweden”, which was participants’ home country. In fact, the average search time for “Sweden” was shorter than for any other country name used in the experiment.

The study reported in this paper is a feasibility study, an initial phase of exploring the VAVS technique. Choosing unassisted visual search as a baseline for comparison was a natural choice for this first step. Further exploration of the technique is planned to compare VAVS with other types of visual search support, such as using text search strings for visually locating objects displayed on the screen. Other possible issues to be explored in future research are as follows:

Augmented reality applications. In augmented reality applications VAVS can be used to help people locate objects of interest in the physical environment. For instance, providing voice input to a wearable system that includes a head up display can help a supermarket customer locate a certain product on a shelf.

Using small screen devices to view large images. Visual search can be especially difficult if the user scans a large image (e.g., a map) using a small screen device, such as a smartphone. A variation of VAVS can be implemented to recognize user voice input and, if it matches an object, which is a part of the large image but not displayed

in the small window, indicate the direction in which the window needs to be scrolled to display the object.

2D sound feedback. A potential problem with VAVS is that in case of very large, complex, and dynamic displays the visual highlighting produced by VAVS could be difficult to detect. A possible solution to this problem is to supplement visual highlighting with a 2D sound signal that would direct user's attention to the general spatial location of the object of interest.

References

1. Andrews, C., Endert, A., North, C.: Space to think: Large, high-resolution displays for sensemaking. In: Proc. CHI 2010, pp. 55–64. ACM Press, New York (2010)
2. Benyon, D., Turner, P., Turner, S.: Designing Interactive Systems: People, Activities, Contexts, Technologies. Addison-Wesley, NY (2005)
3. Bolt, R.A.: “Put-that-there”: Voice and gesture at the graphics interface. In: Proc. of the 7th Annual Conference on Computer Graphics and Interactive Techniques, pp. 262–270. ACM Press, New York (1980)
4. Fabiani, M., Low, K.A., Wee, E., Sable, J.J., Gratton, G.: Reduced Suppression or Labile Memory? Mechanisms of Inefficient Filtering of Irrelevant Information in Older Adults. *J. Cogn. Neurosci.* 18(4), 637–650 (2006)
5. Lau, T., Reed, D.: Speech-activated user interfaces and climbing Mt. Exascale. *Communications of the ACM* 52(6), 10–11 (2009)
6. Manaris, B.: Natural Language Processing: A Human-Computer Interaction Perspective. *Advances in Computers* 47, 2–68 (1998)
7. Nielsen, J.: Voice Interfaces: Assessing the Potential, <http://www.useit.com/alertbox/20030127.html>
8. Volda, S., Podlaseck, M., Kjeldsen, R., Pinhanez, C.: A study on the manipulation of 2D objects in a projector/camera-based augmented reality environment. In: Proc. CHI 2005, pp. 611–620. ACM Press, New York (2005)
9. Ware, C.: Information Visualization. Morgan Kaufmann, Amsterdam (2004)
10. What can I do with Windows Speech recognition?, <http://windows.microsoft.com/en-US/windows7/What-can-I-do-with-Speech-Recognition>

Analysing the Playground: Sensitizing Concepts to Inform Systems That Promote Playful Interaction

Stefan Rennick Egglestone¹, Brendan Walker², Joe Marshall¹,
Steve Benford¹, and Derek McAuley¹

¹ Horizon Digital Economy Research
Sir Colin Campbell Building, University of Nottingham Innovation Park
Triumph Road, Nottingham, NG7 2TU
{sre, jqm, sdb, drm}@cs.nott.ac.uk

² Aerial
258 Globe Road, London, E2 0JD, UK
info@aerial.fm

Abstract. *Playful interaction* is an important topic in HCI research, and there is an ongoing debate about the fundamental principles that underpin playful systems. This paper makes a contribution to this debate by outlining a set of sensitizing concepts which have emerged from an analysis of interaction in the playground; these help explain its appeal to children, and have been selected for their potential to inspire the design of future playful systems. These concepts have emerged from the analysis of material collected during a structured workshop which was organized by the authors, and which was attended by a group of experts. They have also been applied to the design of *Breathless*, a playful interactive system which has recently been deployed by the authors, and which represents an unusual evolution of the playground swing. The paper concludes with a number of reflections inspired by *Breathless*. These have been structured through the use of the concepts as an analytical tool.

Keywords: Playground, playful interaction, sensitizing concepts.

1 Introduction

Playful interaction has been a topic of HCI research for some time, and a variety of authors have provided contributions that have helped shape design. Gaver, for example, has considered aspects of playful interaction that relate to ludic [1] activities such as *exploration*, *invention* and *wonder* [2], whilst Bekker, Sturm and Eggen have emphasised the importance of both *physical* and *social* aspects of playful interaction [3], suggesting a set of values that underpin them, and describing a set of systems that illustrate the application of these values to design. Although some studies have focused on playful interactions between children and computing technology [4], play is an activity that can be beneficial for humans of all ages, and there is a growing strand of research focused on design-led investigations into systems that exploit playful interaction to provide beneficial effects. A significant example is the use of gaming systems to promote whole-body interaction and rehabilitation for individuals

who have acquired physical disability. Here, technological approaches have ranged from immersive virtual-reality systems [5] to commodity gaming consoles such as the Nintendo Wii [6], with the latter becoming regularly used by professional therapists to support individual rehabilitation in both clinical and domestic settings.

Within this broad area of research, the authors have previously pursued a series of design-led investigations that have been inspired by the theme-park environment, and which have resulted in the deployment of systems with which participants have demonstrated a variety of different types of playful interaction. In Fairground: Thrill Laboratory, a number of volunteers were fitted with a telemetry system that captured aspects of their experience on a thrilling ride and transmitted it back to a live audience [7]. Transmitted information included video of the rider's face and quantifications of their physiological responses to the ride; observations and interview analysis have since revealed examples of riders using these systems in a playful way, either by choosing to perform to an unseen audience through the video feed, or by trying to play games with their own physiological response, in the knowledge that it was being observed and interpreted by others. More recently, we have constructed the Broncomatic, a novel, small-scale ride that rotates left and right in response to the direction of a rider's breathing; this has been embedded into a strenuous, competitive game which encourages competitive social interaction between participants [8].

However, in tandem with such design-led interventions, we have recently engaged in research into the principles of systems that promote playful interaction, and a presentation of the outcomes of this research is the central contribution of this paper. Working within a framework that focuses on physical and social aspects of play, we have chosen to shape our contribution by studying the *playground*, a form of entertainment for children which is incredibly popular in most countries, and which has recently emerged as a target of interest within HCI research. Our approach has been to use data collected during a carefully-structured workshop to develop a set of sensitizing concepts that help explain the specific appeal of the playground as a place where children can take part in playful interaction, but which have the potential to inspire the future design of systems that feature playful interaction. We present these concepts in section 5 of this paper, and then illustrate their creative potential in section 6 by discussing their relationship to the design of Breathless, a public event constructed around a novel interactive installation which has been designed by the authors, and which represents a dramatic evolution of the playground swing. Firstly, however, section 2 provides a description of the playground and considers relevant research, and section 3 describes the methodological basis of the work presented in this paper, with a focus on the use of sensitizing concepts to make a contribution to knowledge. Section 4 then provides an overview of the structure and proceedings of the workshop that we organised. After the presentation of sensitizing concepts, we then discuss knowledge that has been gained in relationship to them through an analysis of data collected during Breathless. Finally, the paper concludes with more general reflections inspired by these concepts, and with a brief discussion of the topic of future research structured around them.

2 The Playground

The *playground* is a term that has a variety of meanings across different cultures. For the purposes of this paper, we have constrained the scope of our research by only considering playgrounds that have been designed for children and which are situated outdoors. This differentiates the playground from the park, a space reserved for the leisure of the wider population [9], and from the trend to provide playgrounds that are targeted at adults [10]. Of course, playgrounds may be situated in parks, and adults may have a role in the activities that take place in them, but our focus is on spaces that have been designed with the needs of children in mind. Our rationale is that the design of such spaces has a long and well-documented history, especially in relation to playgrounds intended for urban spaces [11], and, in the UK at least¹, the provision of playgrounds for children has been taken as a shared responsibility of local and national government. This means that there is currently an active community of professionals with a substantial expertise in playground design, as evidenced by the existence of companies that design and construct playgrounds [12]. The playground is also a current topic of research within a variety of academic disciplines, including child development [13] and educational psychology [14]. Given this level of external expertise, the challenge for the authors has been to design a research process that focuses on enabling the integration of prior expertise into HCI research, rather than focusing on a process that generates new knowledge in its own right. Section 3 of this paper describes the methods that we have chosen for this task, and the remainder of this paper considers the results that have been produced by it.

The design of playgrounds for children is also an active area of HCI research, much of which focuses on augmenting playgrounds with interactive features. Much of this research has been constructed around an explicit manifesto of encouraging physical exercise and social interaction, motivated by the need to reduce childhood obesity and to increase social contact (for example, [15]). However, there are also examples of interactive playground equipment that have just been designed to provide interesting new affordances for play [16], rather than changing lifestyles. Within this research effort, Sturm et al [15] have classified potential interventions, using two different set of dimensions. The first relates to the type of intervention that can be made; categories include *installations* of interactive technologies (such as interactive water games [17]), *interactive props* (such as the LEDball [15], an object that lights up in various colours when shaken or rotated), and *interactive surfaces* (such as the ADA floor [18], which can sense people, and respond with different colours of light). A second set of dimensions relates to key design challenges, namely: *social interaction*, *simplicity*, *challenge*, *providing goals* and *feedback*. Using the terminology of these two frameworks, the workshop described in the next section primarily focuses on playground interactions that are orientated around *installations*, rather than *props* or *surfaces*. The deliverables of this workshop are set of sensitizing concepts that suggest approaches to design, and which might therefore be considered complimentary to the design challenges set out by Sturm et al, though they are intended to be more widely applicable than just the playground.

¹ All authors are resident in the UK.

Other examples of HCI or design research in the playground mostly relate to specific interventions that have been made. For example, the Interactive Slide [19] is a large, inflatable slide, intended for use by multiple children, with an interactive surface created using a projector and a video camera, whilst the Interactive Pathway [16] consists of a weight-sensitive pathway which children can stand on, causing various playful actions to occur. An example of an interactive prop is provided by Misund et al [20], who describe an augmented version of a traditional “chase and catch” game. Derakhshan et al [21] describe the use of artificial neural networks to categorise the behaviour of children who are playing on an interactive surface.

3 Research Methods

As noted previously in this paper, a substantial body of academic expertise in relation to the playground already exists. Therefore, in attempting to understand the appeal of the playground, and to generate knowledge that can inform the future development of playful interactive systems, our approach has been to focus on synthesizing and presenting existing knowledge, through the medium of a focused workshop. This was attended by experts from a variety of backgrounds, and its primary purpose has been an examination of the playground from a number of perspectives. The proceedings and structure of this workshop are described in section 4 below.

Material gathered during this workshop has since been analysed by the authors. Informed by a rich tradition of qualitative research, these analyses have been structured to generate a set of *sensitizing concepts*, which are presented in section 5. Sensitizing concepts are analytical constructs that “give the user a general sense of reference” [22] and which can guide attention to particular events or behaviours [23]. They are a tool that is commonly used in areas of research that require the understanding of people and their interactions, including anthropology, sociology and health-care [24]. They are also a tool with a long history of use within research that has impacted on HCI. An example is provided by Crabtree et al [25], who present a set of sensitizing concepts which identify classes of location in the home that are amenable to computation interventions. Such concepts have then inspired the design of systems such as the Drift Table [26] and LINC [27], an electronic family calendar. Sensitizing concepts presented in this paper are intended to explain the success of rides in the playground, in order to inspire future playful systems.

Having presented a set of sensitizing concepts, section 6 illustrates their creative application by considering *Breathless*, a novel installation which represents a dramatic evolution of a playground swing, and which was constructed by the authors. The design and construction of *Breathless* was led by a professional designer, who drew on the sensitizing concepts presented in section 5 for inspiration, and who integrated a variety of other source materials. *Breathless* was experienced by more than 50 participants, and a variety of documentary material was gathered in relation to their experience and the running of the event. An analysis of this material is presented in section 6, where it contributes practical knowledge in relation to the sensitizing concepts outlined in section 5.

4 Workshop Proceedings

As stated in section 2, the focus of our workshop was on understanding the appeal of the static installations that are commonly found in playgrounds. In the UK, these tend to include swings, roundabouts, slides, zip-wires and see-saws. Since playground equipment may vary around the world, representative images of these are shown in figure 1. To ensure that any outputs of this workshop were useful to the research community, we were careful to select a group of participants with relevant professional or academic expertise. These included:

- a professional playground designer from Free Play, a company that focuses on the integration of research into playground design practice [28]
- members of the Learning Sciences Research Institute [29] at the University of Nottingham who had previously conducted research orientated around play
- an experienced ethnographer who had conducted a study of playful interaction
- individuals with relevant expertise in interaction design
- individuals with relevant expertise in HCI
- individuals with relevant expertise in psychology

The workshop was led by a design consultant from Aerial, a company specialising in the design and construction of thrilling experiences [30]. It lasted for a day, and was divided into three sessions, each of which involved work that took place in small groups. Each session began with a short talk by an invited expert, and then involved time to work on a task that was set by the workshop leader. Tasks were designed to draw on the knowledge and experience of participants, both as professionals or academics, but also as adults with personal experience of playgrounds, either as children themselves, or as the parents of children. Tasks were designed to produce tangible outputs, to aid the analysis processes that were planned at the conclusion of the workshop. Tasks were also designed to require both analytical and creative thought processes, with the intention of inspiring participants to think about the playground in a variety of different ways, aiming to produce a level of analysis that was suitably deep to make a useful contribution to HCI. The proceedings of each task were recorded in a variety of ways by the authors of this paper, and have since been analysed. The following sessions and tasks took place in this workshop:

Session One

This session began with a talk by a professional playground designer, who described designs that he had produced, and participatory design methods that he had adopted. Participants of the workshop were then split into groups, and each was allocated two of the pieces of equipment that are featured in figure 1 below. For each, participants were provided with a sheet of representative images as a stimulus, and were asked to draw on their personal and professional experience to produce a description of the appeal of each ride. During this session, participants assembled a number of large paper sheets to which post-it notes had been attached, which were then collected by the workshop organisers. Participants also presented their thoughts to the group. Video-cameras were used to capture these presentations for later analysis.

Session Two

Different groups were formed for this session, and each was assigned a single type of ride to think about, from the selection shown in figure 1. Participants were then asked to design a novel ride inspired by this, using only mechanical technologies. Participants were provided with sheets of paper to sketch their designs on, which were later collected by workshop organisers. Participants were also asked to present their ideas to the whole group, and these were again captured on video.

Session Three

Given the authors previous work on playful technologies which sensed the responses of participants (see section 1), this session began with a talk by an expert, who described a broad range of available sensing technologies, and motivated their use. Participants were then reformed into different groups, and each was given a single type of ride to work on. Participants were then asked to design a novel ride which was inspired by this, but which integrated electronic sensing technologies in some form. For this activity, participants were explicitly told that they could think beyond the boundaries of the playground, and were provided with a selection of materials to construct models of their designs from. As before, group presentations were video-recorded, and photographs of all models constructed by participants were taken. Participants were then given the opportunity to talk about any thoughts that they had in relation to the workshop, which concluded after this session.



Fig. 1. Play equipment commonly found in UK playgrounds. (Clockwise from top left) Swings, slide, zip-wire, roundabout, see-saw.

5 Sensitizing Concepts

After the workshop, the authors discussed the meaning of the data that had been collected, focusing on trying to understand how our participants thought about the

playground. A thematic analysis [31] then led to the sensitizing concepts presented in this section. These have been selected to provide a novel perspective on play in the playground, and for their potential to inspire the future design of playful systems. Each is accompanied by a description drawn from material generated by the workshop, which is intended to support its comprehension by the reader.

Danger and Fear Are Important Playground Experiences

Many of our participants spoke of exhilarating playground experiences that involved them performing dangerous activities, or being scared of the situations they had got themselves into. Participants talked about using swings as launchers, allowing them to fly through the air and land on the ground, sometimes causing themselves injury, but more normally just enjoying the thrill and pleasure of using their bodies to provide such momentary experiences. Participants also talked about spinning too quickly on roundabouts, worrying about flying off, but being exhilarated by the physical challenge of hanging on despite centrifugal forces. The professional playground designer who introduced the first session talked about expensive playgrounds which had been “KFCed” – Kitted, Fitted and Carpetted for safety – and which were almost always underused because of this. Participants also talked about the social fears that could be created by interactions between the different age groups that might use a playground, but focused on the importance of such situations for learning how to interact in a social environment. In particular, playing with the older children on a ride for the first time was seen as a scary but important social barrier to overcome by some participants, and an important part of the process of growing up.

Loss of Control Is a Key Component of Playground Installations

Our participants considered that some of the exhilaration described in relation to the previous concept is provided by the loss of control that is inherent in many rides. This is obvious on an installation such as a zip wire, where, after launching, the rider allows themselves to be controlled by gravity, potentially leading to an exciting ride experience whose duration can barely be influenced. Loss of control can be part of the experience of taking part in a swing, since a rider can be pushed higher and higher by someone standing on the ground, allowing them the thrill of exceeding their own comfort zone in a way which they may not be able to manage on their own. The sharing of control is a key part of the experience of a see-saw, in which control oscillates in a rhythmic way between the two participants, both of whom have some control over, and some responsibility for, the others safety during this cycle. Participants also considered loss of control to be an important part of the roundabout experience, as a rider who had boarded a roundabout could then be spun faster and faster by those standing on the ground, causing them to have to hang, fall off or take the risk of jumping off and potentially injuring themselves.

Play Often Involves the Re-appropriation of Installations in Unexpected Ways

Participants talked about re-appropriation as a key issue, and described examples of it leading to enjoyable forms of play. Participants talked about the fun of using a swing in ways for which it was not designed, either by trying to swing left and right (rather than forwards or backwards), by trying to balance on it in unusual ways, by getting several people to ride the swing at once, or by using it as a launcher. One participant

described a game in which he lay on the ground underneath a swing seat, which was then thrown towards his head. The chains attached to the seat would restrain it, but the play of the game involved flinching as little as possible as it was thrown. There were many other examples of appropriation that were given during the workshop, including the re-appropriation of equipment by older groups of children for “social loafing”, therefore allowing them to exert control over access to it by younger children. A form of re-appropriation that was emphasised by the professional playground designers was that of fantasy – i.e. children using their imagination to think of the playground in different ways, such as a magical forest. Interestingly, the slide as a piece of playground equipment was criticised by a number of participants as having little potential for re-appropriation, although children could still climb backwards up it.

Installations That Cater to Different Levels of Skill Can Support Long-Term Play

Many playgrounds are used by children with very different ages, and different (and often rapidly developing) levels of physical ability. Playgrounds, although simple, often maintain their level of interest across many years of a child’s life, and our participants considered this to be a beneficial product of the very different levels of skill that could be used on a variety of items of equipment in a playground. For example, a very young child might just enjoy sitting in a swing, under the control of a parent, whilst an older child might enjoy learning to control the movements of their body, in order to swing higher than before. A very developed child might then enjoy the physical challenge of standing on a moving swing. Pushing others on swings is also a playful activity that children of many ages and levels of ability could engage in. As well as being playful and fun, such interactions with playgrounds have the potential to assist the learning of new motor skills, with positive benefits for the development of children. Many of the re-appropriations of playground equipment that participants considered also involved a high level of skill for mastery, and developing a new skill was seen as part of the fun of taking part in these activities.

Different Designs of Rides Offer Very Different Throughputs

Playgrounds are rarely managed environments, and the throughput inherent to particular rides has an impact on the experience of using them. For example, a slide has a fairly clear trajectory [32] in which a participant starts at the top and slides down, unless it is re-appropriated by children who wish to climb back up. Such a trajectory is likely to maintain a high throughput, as the duration of each experience is limited by the length of the slide. In contrast, a roundabout or swing can be used in an open-ended way, and participants may only get off when tired. Such open-ended trajectories may allow participants to experiment with their experience in a playful way, but may cause frustration for other children who are denied access to a quickly moving roundabout or an occupied swing for a long period of time.

Spectatorship and Performance Are an Important Part of the Experience of Play

Many discussions of playground equipment focused on the rider, but spectatorship was also seen as an important role in play, albeit often a transitory one. For example, a child who was trying to push themselves higher and higher on a swing might gain even more enjoyment and motivation if being watched by a peer or a parent, and a

spectator who was a child might be inspired to gain a new skill by watching others in the playground performing on rides. Spectatorship is also part of the social interaction in a playground, and there could be some tension between a child who is controlling a piece of equipment, and who wants to carry on using it, and a child who is watching the ride and wanting to use it. Equally, the control of a ride by a social group could be seen as a visible display of power, with understanding how to subvert such power on the part of a spectator being an important social lesson.

Table 1. Summary of sensitizing concepts from the playground, with key elements highlighted in first column

<i>Danger and fear</i>	The excitement provided by a level of danger and fear is important to the success of playgrounds.
<i>Control</i>	Losing control and being pushed past your comfort level is a key component of some playground installations.
<i>Re-appropriation</i>	As part of imaginary and exploratory play, installations are often re-appropriated for purposes that they were not designed for.
<i>Skill</i>	Catering for different levels of skill and development is important for supporting long-term play.
<i>Throughput</i>	Different designs of rides offer very different throughputs, which affects the social use of these rides.
<i>Spectatorship and performance</i>	Watching others on rides, and demonstrating or showing off when on the ride are important parts of playground interaction.

6 Knowledge Gained through Study of a Novel System

Section 5 has described a set of sensitizing concepts developed from a workshop themed around the playground, which are intended to provide inspiration for the design of systems that promote physical and social interaction. In keeping with the traditions of this method, we would expect these concepts to be modified, challenged or re-appropriated by ourselves and others as they are integrated into future work. As part of this process, this paper now presents an account of our initial work in applying them, as part of a process which led to *Breathless*, an installation which was assembled in the autumn of 2010, and which was publicly deployed on a single evening. We begin this section with a description of *Breathless*, and then present an analysis of material collected during its deployment, which is used to highlight the relationship between this installation and the set of concepts presented in section 5. This paper then concludes with a broader discussion of the practical knowledge that has been gained in applying these concepts, and by considering future work in relation to the development of our sensitizing concepts.

6.1 Overview of Breathless

Breathless is an experience constructed around a very large rope swing, which is driven by a powerful electric wheelchair motor controlled by a computer. Figure 2

illustrates the use of this motor to drive a horizontal rope, which in turn transfers energy directly into the rope swing, and therefore affects its movements. For its first deployment, the swing was suspended from the covered roof of an outdoor exhibition space. The horizontal rope made use of two pulleys attached to pillars.

In keeping with the Broncomatic [8] which was described in the introduction to this paper, in *Breathless*, the movement of the horizontal rope is influenced by the breathing of the rider. In particular, breathing in causes the rope to move one way, and breathing out causes it to move the other way. As such, breathing in synchronization with the natural oscillations of the swing causes it to go higher, whilst breathing out of synchronisation caused it to judder and go lower. The length of the vertical rope that connects the swing seat to the loading point in the ceiling of the space is calculated to create a resonant frequency of 5 Hz (i.e. 12 breaths per minute, which is a natural rate of breathing for humans at rest). The idea is to create an experience that responds strongly to breathing that is in harmony with it, but which feels uncomfortable given breathing that is not in harmony.

Breathless uses a modified gas mask to monitor the breathing of participants, through a custom-designed filter which is screwed into a port in the mask. An image of a participant wearing one of the masks is shown on the left of figure 3. This features a participant who is lying in a custom-made cradle which replaces the seat or knot more commonly found at the bottom of a rope swing. This is designed to make it difficult for a participant to transfer energy into the swing through movements of their body, ensuring that the swing is controlled through the breathing of participants alone.

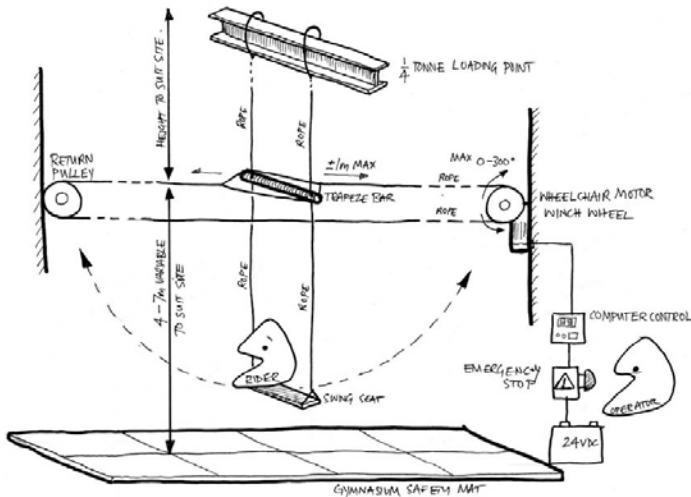


Fig. 2. Illustration of mechanism of rope swing

In addition to the unique form of interaction provided by the mechanism described above, the experience of taking part in *Breathless* is heightened by the structure and theming of the event, which together create an intense experience with a strong sense



Fig. 3. Left: Participant riding the swing whilst wearing the gas mask. Right: Participant in seat, waiting for next rider to mount.

of immersion, but which still allowed for playful interaction between participants. To accentuate the atmosphere of the experience, the event took place after dark. We also used a wireless microphone attached into the gas mask to collect the sound of participants breathing, and used a powerful amplifier to fill the event space with this sound.

A final element of the event relates to the sensitizing concept *loss of control* which emerged from the workshop described above. Rather than just allowing participants to control their own experience on the ride, every participant in *Breathless* actually moves through three different pre-defined roles, allowing them to experience different aspects of control. In general, movement through *Breathless* is tightly choreographed by event staff, and works as follows:

1. Participants join a queue take part, and arrive at a desk, where they are fitted with an appropriately-sized gas mask, fitted with a breathing sensor, which has been sterilised if previously used
2. This breathing sensor is then activated, and a visualisation of the participant's breathing appears on a very large projection screen which is visible to all event attendees. This projection screen integrates visualisations of live breathing data being collected from all participants currently wearing gas masks
3. The participant is escorted to a position from which they can observe the operation of the ride (which at this point is being controlled by other participants)
4. The participant is then assisted to mount the cradle, and their movements are controlled for a while by the participant who has just vacated it (and who is now sitting in a chair next to it, as shown in the photograph on the right of figure 3)
5. At some point, control is then switched to the participant who is in the cradle, whose breathing then influences their own movements on the ride for a while.
6. Finally, the participant in the cradle dismounts, and takes the now-vacant position in this seat to control the next rider.

6.2 Collection and Analysis of Material during *Breathless*

Breathless was designed and constructed over a two-week period, and opened to the public on a single evening, where it was experienced by over 50 participants.

Throughout the process of conducting *Breathless*, we collected a variety of material for later analysis. These included semi-structured interviews with twenty-four participants, ambient video recorded from a number of viewpoints within the exhibition space, and regular interviews with designers during the two-week design process. Much of this material has since been transcribed and analyzed. We now present a number of elements of this analysis, structured by the sensitizing concepts introduced in section 5. This interview material has provided insights that will guide the future of development of *Breathless*, and other similar experiences. We return to it in the final section of this paper, where we discuss wider issues inherent in designing novel playful systems.

6.2.1 Danger and Fear

A number of elements in *Breathless* were selected for their potential to create fearful, intense experiences, but with no real threat of danger to our participants. For example, our use of gas masks seemed to create the potential for uncomfortable feelings of claustrophobia, and we wondered whether the height that the swing seat could rise to might be emotionally uncomfortable for riders, especially given the amount of momentum which is inherent in the system. We asked participants about these issues in our interviews, and have since collated responses.

Although some participants were clearly made uncomfortable by wearing a gas mask, an analysis of responses suggests that there are actually very subtle differences in perception and past experience between participants that influenced their experience, raising an interesting issue of how to design around the concept of danger and fear in the future. For example, one participant had very poor eye-sight, and had to remove his glasses to put the gas mask on – he then described the resultant hazy view that he had from the swing as being “nerve-wracking”, but enjoyable because of this intensity. In contrast, a number of participants who were experienced scuba divers described how wearing the gas mask actually felt enjoyable, because it reminded them of diving. However, one participant, who had been diving once and who had hated it, described how the gas mask reminded her of this unpleasant experience. In terms of the height that the swing reached, most participants stated that this was not sufficient to scare them, and that they would have liked it to have gone higher. However, one participant reported being so scared of how high that he was going that he felt like holding his breath, to make sure that it did not go any higher. Finally, although many participants reported feeling claustrophobic when they first put the mask on, a number described how, once they relaxed into this experience, they actually started enjoying the experience of wearing it. In relation to this, one participant described how the mask made her feel like she was in a space of her own, whilst another participant described the experience as like being alone in a bubble, and implying that this was an interesting and comfortable situation to be in.

6.2.2 Control

Clearly control is a key part of the experience of *Breathless*, with participants being in control when they first mount the swing, then losing control at some point in the experience, and then being given the chance to exert control over another participant after they have vacated it. One participant described this final phase as being “powerful”, especially given that the participant he was controlling was his friend.

Another participant described *Breathless* as a very playful experience, because of the option of either “making the other person more thrilled by their experience” or “teasing them to experience something a bit more”. Most people described wanting to give other participants a nice ride, although one participant described wanting to create “really rough high frequencies and fluctuations so the person would hopefully be a bit more freaked out”. One participant described trying to be really careful when controlling the ride of a participant who appeared scared, whilst other participants described the added interest of trying to control someone that they already had a relationship with, such as friends, brothers or sisters. When being controlled by others, one participant reported a bad experience, given by a person who “didn’t really cope with the concept ... it was more like ‘bump, bump, bump’, it was more like a wobbly chair”. However, others described really pleasant experiences: “it was good, because he really had the knack of it, like he was doing it really smoothly”. Interestingly, some participants tried to communicate with the person who was controlling them, but found this difficult, because the gas mask restricted communication. In addition, because control of the swing is somewhat indirect, a number of participants reported never fully feeling sure about when control shifted from the person sitting on the chair to them, though this lack of certainty seems to have been interesting in its own right.

6.2.3 Re-appropriation

Participants were only given a few minutes in each of the three roles in *Breathless*, giving little time for the kinds of re-appropriation described in section 5. However, interview comments by certain participants have helped us to think about re-appropriation in more depth; in particular a number of interviewees described not being able to work out what the purpose of the *Breathless* setup was, and therefore having to experiment for a while before working out how to use it (and enjoying this experimental part of the experience). These comments remind us of work on ambiguity in design (for example, [33]). We wonder whether designs that are more ambiguous are more amenable to re-appropriation – or, alternatively, whether the re-appropriation of designs that have an apparently clear purpose is more enjoyable.

6.2.4 Skill

From our own experimentation, we expected participants to take some time to learn how to control the swing in *Breathless*, and interview material provides some detail about the challenge of this process. An analysis of this material suggests that some participants sometimes drew on previous experience to learn how to control the swing, alongside experimentation with their breathing and its effects on the system. For example, one participant described how martial arts training had led him to consider the ratio between his in-breath and his out-breath, which led him to experiment with varying this ratio whilst riding the swing. Other participants described experimenting with the smoothness of their breathing, with the point in the swing-cycle that they breathed at, and with breathing with either their mouth or their nose. One participant stated that he had a bad cold, and described how this made it hard to control the ride. Other participants described having a lack of an intrinsic ability to keep a rhythm, and suggested that this made it harder to get the swing moving higher. Finally, some participants reported that it was much easier to judge

the movement of the swing when they were sitting beside it, controlling someone else. In comparison, when riding the swing, they had little vision, so had to try and sense the position of their bodies, by paying attention to the weight that they were putting on the swing-seat, their sense of balance and the visceral feelings of acceleration that were induced in their stomach.

6.2.5 Throughput

For the designers of *Breathless*, throughput was a key issue when preparing for the public event. A large number of participants were expected, and there was no clear incentive for riders to dismount the swing, and to give way to others. The movement of participants through three roles was a clear response to this issue, and movement between these roles was created in a variety of ways. Primarily, event staff were in charge of the electronic linkage between breathing sensor and ride movement, and could disconnect this when a participant had experienced a reasonable amount of time on the swing. Other event staff were then on hand to move participants through the positions that related to roles. Such a high level of choreography contrasts strongly with the situation in a typical playground, in which throughput is managed through social interaction; clearly, therefore, although choreography is one tactic for maintaining throughput, other tactics may be worth exploring.

6.2.6 Spectatorship

A number of participants reported how the experience of spectating *Breathless* influenced their experience of taking part. For example, one participant reported watching the visualization of her breathing that was displayed on the projection screen, and learning how the trace responded to in-breaths and out-breaths. Another described being made to feel nervous by the spectacle of a person in a gas mask, swinging backwards and forwards in the dark. One participant noticed that her breathing appeared very different to other participants, and practiced making it more similar – which then helped her on the ride. Finally, another participant reported swearing by mistake, and then realizing that this had been heard by the whole of the audience – which therefore made her feel self-conscious.

7 Discussion – Lessons Learned, and Future Work

In this paper, we have presented a set of sensitizing concepts derived from an analysis of the playground, and illustrated their creative application to *Breathless*, an installation which represents a dramatic evolution of the playground swing. In this section, we now present some brief reflections on the knowledge that has been gained through this process, structured through two core themes – accounting for individuality and considering trade-offs between concepts. We also provide a commentary considering the relationship between inclusive design and playful interaction.

7.1 Accounting for Individuality

A key observation in relation to the research material gathered through *Breathless* is the impact of participant history and personality on the nature of their experience in

this event. To give two examples: participants who had previously had pleasurable experiences whilst scuba diving seemed more likely to enjoy the isolating feeling of wearing the gas-mask during the event, and participants who had worn gas-masks at school (possibly through drills designed to practice response to chemical attacks) seemed more likely to at least be comfortable with wearing the mask. There also seemed to be some differences between participants in relation to their perceptions of the height that the swing rose to, with some participants being scared by this, but with other participants wanting a more extreme experience than we could provide. Collectively, these observations hint at the challenge of designing playful experiences that are enjoyable and beneficial to those that take part in them. In relation to these observations, two opposing tactics suggest themselves:

Aim for universality: This tactic, which may be closest to traditional playground design, is to aim for experiences that are as universal as possible. The concepts presented above provide a resource for working within this tactic, but it must present a significant challenge due to the nature of individuality. Universal designs might be relatively generic (like most playground equipment). Designers might also consider approaches such as focusing on installations that can be dynamically adapted to the personality and abilities of their users, to maximize the range of individuals for whom they are relevant.

Design for specific groups: Abandoning universality allows the freedom to design for specific groups of individuals. In relation to *Breathless*, for example, we could imagine a variant that provided a far more thrilling experience for those who would appreciate it (i.e. an approach which emphasizes *danger and fear*), or a second variant that was more focused on supporting *re-appropriation*. This tactic raises an issue of how to identify groups that are interesting to design playful systems for. It also raises issues such as how to profile participants and allocate them to groups.

7.2 Considering Trade-Offs between Concepts

A second issue that is highlighted by material collected during *Breathless* is the potential trade-offs that might have to be made when using the concepts presented in this paper as a resource for design. In *Breathless*, for example, a clear trade-off had to be made between allowing participants sufficient time on the swing to start to think about re-appropriation and maintaining a sufficient throughput to give all attendees at the event a chance of getting involved. This relates to the nature of the event, which only ran for one evening, and which presented a novel experience which many attendees wished to sample. Playgrounds negotiate this trade-off differently – by often being continuously accessible (and therefore reducing contention) and by relying upon human interaction to resolve contention. Equally, however, contention in the playground can lead to some users being denied access to equipment.

Other trade-offs undoubtedly exist between these concepts, and exploring them seems to be an interesting direction for research. When a trade-off has to be made, it may be that investigations into the available tactics within that trade-off could provide benefits.

7.3 Inclusive Design in Relation to Playful Systems

Finally, we wanted to acknowledge the potential importance of inclusive design and playful systems, in relation to the concepts that we have outlined above, and especially in relation to play that encourages physical and social interaction. In particular, we wanted to highlight issues around the concept of danger and fear. Even in today's risk averse societies, playgrounds still offer an opportunity for the able-bodied to experience the pleasures of flirting with danger, albeit with some safeguards to help avoid injury, such as play surfaces to fall onto that are soft in comparison to older materials such as concrete or gravel. However, a number of authors have argued that, when seeking to design for individuals with disabilities, there is often an excessive emphasis on safety, which carries a risk of denying worthwhile experiences around danger and fear to these individuals [34]. When seeking to design within the set of concepts presented in this paper, there is therefore an interesting challenge of how to ensure that as wide a group of individuals as possible are given the opportunity of gaining benefit from playful systems that embed these concepts. In particular, we might pose the question of how to design playful systems that are amenable to interaction for individuals with disabilities, but which still allow for creative re-appropriation, and which still allow for controlled elements of danger, for those participants that wish to take risks.

8 Conclusions

Playgrounds provide a relatively universal experience which involves physical and social interaction, and which encourage play. Our analysis of a workshop focused on the playground has provided significant inspiration for the design of *Breathless*, a novel interactive experience, and we hope that it will provide inspiration for others. We can imagine future work that involves novel design, and which draws on these concepts. We can also imagine future work that uses them as an analytical framework when considering existing design.

Acknowledgments. This work was supported by the Horizon Digital Economy Hub (EP/G065802/1).

References

1. Huizinga, J.: *Homo Ludens: A study of the play element in culture*. The Beacon Press, Boston (1950)
2. Gaver, W.: Designing for Homo Ludens. *I3 Magazine* (12) (2002)
3. Bekker, T., Sturm, J., Eggen, B.: Designing playful interaction for social interaction and physical play. *Personal and Ubiquitous Computing* 14, 385–396 (2010)
4. Nielsen, R., Fritsch, J., Halskov, K., Brynskov, M.: Out of the box: exploring the richness of children's use of an interactive table. In: 8th International Conference on Interaction Design and Children (2009)
5. Kizony, R., Katz, N., Weiss, P.: Adapting an immersive virtual reality system for rehabilitation. *Journal of Visualization and Computer Animation* 14, 261–268 (2003)

6. Online guide to the use of the Nintendo Wii in physical rehabilitation, <http://www.wiihabilitation.co.uk/>
7. Walker, B., Schnädelbach, H., Rennick Egglestone, S., Clarke, A., Ng, M., Wright, M., Rodden, T., Benford, S., French, A.: Augmenting amusement rides with telemetry. In: International Conference on Advances in Computer Entertainment Technology (2007)
8. Marshall, J., Rowland, D., Rennick Egglestone, S., Benford, S., Walker, B., McAuley, D.: Breath control of amusement rides. In: 29th International Conference on Human Factors in Computer Systems (2011)
9. Jones, K., Wills, J.: The invention of the park: from the Garden of Eden to Disney's Magic Kingdom. Polity Press, Cambridge (2005)
10. Internet article on adult playgrounds, <http://www.independent.co.uk/news/uk/home-news/eastbourne-to-provide-playground-for-the-elderly-1029067.html>
11. Solomon, S.: American playgrounds: Revitalizing community spaces. UPNE, New England (2005)
12. Playground designers' website, <http://www.schoolplaygrounddesigners.co.uk/>
13. Ladd, G., Price, J., Hart, C.: Predicting preschoolers' peer status from their playground behaviors. *Child Development* 59(4), 986–992 (1988)
14. Barbour, A.: The impact of playground design on the play behaviors of children with different levels of physical competence. *Early Childhood Research Quarterly* 14(1), 75–98 (1999)
15. Sturm, J., Bekker, T., Groenendaal, B., Wesselink, R., Eggen, B.: Key issues for the successful design of an intelligent interactive playground. In: 7th International Conference on Interaction Design for Children (2008)
16. Seitnger, S., Sylvan, E., Zuckerman, O., Popovic, M., Zuckerman, O.: A new playground experience: going digital? In: Alt. Chi at the 24th Annual SIGCHI Conference on Human Factors in Computing Systems (2006)
17. Pares, N., Durany, J., Carreras, A.: Massive flux design for an interactive water installation: WATER GAMES. In: 2nd International Conference on Advances in Computer Entertainment Technology (2005)
18. Delbruck, T., Whatley, A., Douglas, R., Eng, K., Hepp, K., Verschure, P.: A tactile luminous floor for an interactive autonomous space. *Robotics and Autonomous Systems* 55, 433–443 (2007)
19. Soler-Adillon, J., Pares, N.: Interactive slide: An interactive playground to promote physical activity and socialization of children. In: Alt. Chi at the 27th Annual SIGCHI Conference on Human Factors in Computing Systems (2009)
20. Misund, G., Holone, H., Karlsen, J., Tolsby, H.: Chase and catch – simple as that? Old-fashioned fun of traditional playground games revitalized with location-aware mobile phones. In: 6th International Conference on Advances in Computer Entertainment (2009)
21. Derakhshan, A., Hammer, F., Hautop Lund, H.: Adapting playgrounds for children's play using ambient playware. In: IEEE/RJS International Conference on Intelligent Robots and Systems (2006)
22. Blumer, H.: What is wrong with social theory? *American Sociological Review* 18, 3–10 (1954)
23. Holloway, I.: Basic concepts for qualitative research. Wiley-Blackwell (1997)
24. Bowen, G.: Grounded theory and sensitizing concepts. *International Journal of Qualitative Methods* 5(3) (2006)

25. Crabtree, A., Rodden, T., Hemmings, T., Benford, S.: Finding a place for Ubicomp in the home. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) *UbiComp 2003*. LNCS, vol. 2864, pp. 208–226. Springer, Heidelberg (2003)
26. Gaver, W., Bowers, J., Boucher, A., Gellerson, H., Pennington, S., Schmidt, A., Steed, A., Villars, N., Walker, B.: The drift table: designing for ludic engagement. In: 22nd Annual SIGCHI Conference on Human Factors in Computing Systems (2004)
27. Neustaedter, C., Bernheim Brush, A.: “LINC-ing” the family: the participatory design of an inkable family calendar. In: 24th Annual SIGCHI Conference on Human Factors in Computing Systems (2006)
28. Free Play, <http://www.freeplaydesigns.com/>
29. Learning Science Research Institute, <http://www.lsri.nottingham.ac.uk/>
30. Aerial, <http://www.aerial.fm/>
31. Boyatzis, R.: *Transforming qualitative information: thematic analysis and code development*, 1st edn. Sage publications, Thousand Oaks (1998)
32. Benford, S., Giannachi, G., Koleva, B., Rodden, T.: Temporal trajectories in shared interactive narratives. In: 26th Annual SIGCHI Conference on Human Factors in Computing systems
33. Gaver, WW., Beaver, J., Benford, S.: Ambiguity as a resource for design. In: 21st Annual SIGCHI Conference on Human Factors in Computing Systems (2003)
34. Inclusive design case studies, <http://www.designcouncil.info/inclusivedesignresource/benwilson/intro.html>

Comparative Feedback in the Street: Exposing Residential Energy Consumption on House Façades

Andrew Vande Moere¹, Martin Tomitsch², Monika Hoinkis³,
Elmar Trefz⁴, Silje Johansen², and Allison Jones²

¹ Design Lab, K.U. Leuven, Belgium

² Design Lab, University of Sydney, Australia

³ University of Applied Sciences Potsdam, Germany

⁴ DAB, UTS, Australia

andrew.vandemoere@asro.kuleuven.be,
martin.tomitsch@sydney.edu.au, hoinkis@fh-potsdam.de,
elmar.trefz@student.uts.edu.au, sjoh1918@uni.sydney.edu.au,
allisonjones@uni.sydney.edu.au

Abstract. This study investigates the impact of revealing the changes in daily residential energy consumption of individual households on their respective house façades. While energy feedback devices are now commercially available, still little is known about the potential of making such private information publicly available in order to encourage various forms of social involvement, such as peer pressure or healthy competition. This paper reports on the design rationale of a custom-made chalkboard that conveys different visualizations of household energy consumption, which were updated daily by hand. An in-situ, between-subject study was conducted during which the effects of such a public display were compared with two different control groups over a total period of 7 weeks. The competitive aspects of the public display led to more sustained behavior change and more effective energy conservation, as some graphical depictions such as a historical line graph raised awareness about consumption behavior, and the public character of the display prompted discussions in the wider community. The paper concludes with several considerations for the design of public displays, and of household energy consumption in particular.

Keywords: persuasive computing, public display, urban screen, visualization, sustainability, interaction design, urban computing.

1 Introduction

Residential energy consumption is estimated to account for 11% of energy consumption worldwide and is estimated to grow between 0.6 and 2.4% per year [1]. Electricity currently is the highest contributor to residential energy consumption, and it has been estimated that electricity will account for nearly 60% of overall residential energy growth over the next 20 years [1]. As Australia receives approximately 77% of its electricity demands from burning coal, domestic electrical appliances will be to blame for the largest production of greenhouse gases from all residential energy consumption. While people are becoming increasingly aware of the ongoing “Climate

Crisis” [2], they are rarely aware of how their daily activities contribute to greenhouse gas emissions [3, 4, 5]. As a result, the majority of interactions with energy-consuming appliances occur without conscious consideration of their environmental impact [6]. Recent advances in wireless sensor technology, smart metering and electronic displays present an opportunity for advanced forms of behavioral feedback, such as screens that display appropriate, real-time information of the actual energy consumption within the context of everyday life. Commercially available energy usage feedback displays have now become affordable, which typically convey the real-time energy consumption through displaying numerical data, such as kilowatts or financial costs per hour. The idea of feedback is not new, and various forms of energy usage feedback have already been investigated, demonstrating how it has indeed the potential to promote energy conservation to the order of 5-10% in common households [7] or office environments [8]. Moreover, modern communication technology has now the power to act as facilitator for motivating behavioral change through social cues [9], although such interventions typically require the development of novel sensors, visualizations, interfaces or interactions [10]. Several strategies have already been proposed for designing persuasive technologies [11], such as the requirement to converge motivation, ability and trigger at the same moment in time [12]. However, little is known about how energy usage feedback could benefit from the integration of other persuasive means that reach beyond the immediate display of private information, such as inducing forms of social pressure, competition or cooperation by externalizing the feedback beyond the end user. Our study investigates whether turning behavioral information that is normally kept well-hidden and private, explicitly public and even comparable to those of others, can effectively augment persuasive feedback. More concretely, we describe the design, implementation and evaluation of a new public, urban display that presents the individual energy consumption performances of families on their respective house façades.

2 Background

Early studies on the feedback of energy consumption were typically carried out by psychologists, who mainly focused on the reinforcement of certain behaviors through direct intervention. More recently, academic research has shifted to more qualitative studies, in order to understand how people respond to different forms of feedback methods. Research about the performance of environmental feedback is still relatively limited, but ranges from informed billing, smart meters, direct feedback displays [13, 14], numerical read-outs [15], bar graph charts [3], highly detailed information dashboards [16], or ambient cues [17], such as a lights that change color [18]. Several design-based research projects have instead focused on reinforcing the persuasive message of energy usage feedback by including qualities of joy, tangibility or ambiguity, resulting in various design-led projects that are more speculative and risk-taking: some projects dealt with feedback in public space [19], while others proposed novel interfaces for the home [20], or compared graphical forms of feedback [21]. As the potential of persuasive visualization is still relatively unexplored [22], academic research in this realm is still increasing (e.g. the “BeAware” project [23]).

To design successful feedback mechanisms that intent to change human behavior, it is necessary to understand what motivates people. For instance, most commercially available feedback displays rely on an *intrinsic* rational-economic model, which assumes that people can be encouraged to change behavior by the prospect of saving money [24]. Most material incentives and persuasive prompts also have the potential to trigger behavior change [25], but tend to become less effective once the novelty declines or the incentives have been removed [26]. Techniques based on *extrinsic* forms of motivation, such as social reinforcement, can help to discover more intrinsic motivations and even lead to sustained change [24].

In particular, providing *comparative* feedback may have a positive effect on behavior change by triggering feelings of competition, social comparison or social pressure [26]. However, the usefulness of comparative feedback, which contrasts the consumption of multiple people against each other, is still relatively contested. While some early field studies have shown positive effects [3, 27], there has been some evidence that people tend to express concern about the apparent validity of the comparison groups [5]. Other studies also demonstrated that while high and medium consumers conserved energy, some low consumers tended to increase their consumption, as they felt less encouraged when noticing the higher apparent average usage [27]. The study of *Wattsup*, a Facebook application for energy monitoring, indicated that competitive feedback is more enjoyable and more effective compared to individual feedback [28]. *EnergyWiz*, a mobile phone application, employed different forms of social comparison in order to gain insights into the design of comparative feedback [29]. With the recent advances in networked tracking devices, public comparison of behavior for cooperative or competitive purposes has become possible. For instance, the commercially available Apple iPod/Nike+ pedometer promotes competitive running between athletes, even when they are physically separated. Several initiatives based on social media have specifically focused on externalizing energy consumption, ranging from dedicated websites such as *Make Me Sustainable* [30] and *Carbon Rally* [31], through embeddable widgets such as the *Google Powermeter* [32], to augmented energy bills [33]. Most comparative feedback thus still occurs in the online realm, shifting the context of feedback away from physical reality, in which energy usage actually occurs.

In this paper, we propose a new form of *urban display* [34]: a public display that represents information that is relevant and contextualized to its immediate surroundings. While electronic displays are becoming increasingly ubiquitous in today's public space, the majority serves mainly commercial, artistic or entertainment purposes. In contrast, our project foresees a future in which public displays offer information that is socially relevant, and encourages local support or cohesion.

3 Design

The general aim of the study was to measure the persuasive effectiveness of public feedback. We decided to apply our research to the issue of energy consumption, because energy usage is relatively simple to relate to behavior and changes can be detected accurately. Other behavioral data such as from water or gas consumption are less detailed, and relatively cumbersome and expensive to capture. Due to the encompassing context of sustainability, the design constraints became manifold and

complex, as, ideally, any (public) feedback display of sustainable behavior should be: 1) *Sustainable* in and by itself, as the construction and maintenance of the system should not negate the intentions it attempts to promote. This constraint proves to be extremely restrictive, as it means the consumption of materials or energy should be minimized, or even completely avoided; 2) *Affordable*, as supporting sustainable behavior should only induce costs that can be earned back over a reasonable amount of time; 3) Respectful to *privacy*, as any communication of behavioral information in public might introduce unintended, yet significant, risks; 4) *Intuitive*, as any feedback should be easy to understand and enjoyable to use, even for occasional passers-by; 5) *Robust*, including a resistance to uncontrollable aspects such as severe weather conditions, vandalism or carelessness when located in the harsh reality of semi-public space; 6) *Aesthetic*, as public feedback should be presented in a form that is acceptable by all its stakeholders while being unobtrusive to everyday activities; 7) *Updatable*, as feedback should be provided shortly after the occurrence of behavior, allowing people to immediately link the impact of their actions; and, 8) *Persuasive*, as feedback should also reveal the meaning behind the information shown, and highlight specific relevant aspects that can more deeply influence intrinsic motivation.

3.1 Public Display Design

The public feedback display was developed through a design-oriented research approach. This involved the execution of successive designs on the basis of iteratively refined constraints and requirements, and the production and evaluation of various low-fidelity sketches, mockups and working prototypes. We decided on the house façade as the ideal location for conveying feedback in a public context: a large sign mounted on a house façade intuitively refers back to the inhabitants that cause the behavior, while it also ensures uninterrupted visibility to neighbors or passers-by. We investigated existing forms of social competition that typically take place in the semi-public domain of the house façade, porch or front garden, including gardening and vegetable competitions, garden gnome collections, Christmas decorations, and the like (see Figure 1, Left). This approach further convinced us of choosing a ‘non-electronic’ display medium, which seemed to be better suited to the physical language of the street; avoided complex and expensive construction issues (e.g. wiring, casing); and solved most of the previously listed (public) feedback design requirements. While searching for a physical material that can display and be easily updated, we arrived at *chalkboards*. Chalkboard surfaces are associated with many positive qualities in terms of our everyday experience: they naturally imply some sort of dynamism and time-variance, exemplified by signs that advertise the changing menus of cafés or restaurants; they convey a warm, handcrafted charm through the often clumsy but sympathetic handwriting and low-fidelity chalk aesthetic; they are surrounded by an aura of playfulness, reminding people of their childhood with activities like doodling, street graffiti, doorstep games, and so on; and they infer a personal and small-scale context, which many of us unconsciously use to, for instance, distinguish the neighborhood coffee bar from the large, international coffee chain franchise. Chalkboards have already proven to be an ideal platform to convey dynamic information, such as game scores, supermarket prices, or the water temperature in the public swimming pool. Even more, we expected that the manual act of updating a chalk-

board could open up the opportunity for people to casually interact with the writer, so that a ‘data update’ has the potential to grow into a true ‘social event’, contrasting the unnoticeable millisecond blip of an electronic display.



Fig. 1. Left: Existing forms of social competition in the semi-public domain of the house façade and front garden. Right: A neighborhood map highlighting the participating households.

We deliberately chose a neighborhood in the city of Sydney that is characterized by a distinct building typology of so-called ‘terrace houses’ and enjoys high pedestrian traffic. The strong similarity in dimension and geometrical layout of terrace house façades allowed us to create displays that were mostly identical, thereby reducing construction complexity and avoiding potential bias in visual perception. However, as in many other cities, altering the appearance of an exterior façade requires explicit approval from the city council, a laborious and time-consuming process that would significantly complicate our intended time planning. An elegant solution was found by replacing our initial concept of traditional wooden blackboards in favor of light-weight twin wall polypropylene sheets. The explicit temporary character of this material allowed us to install the displays with the exemption of the so-called ‘State Environmental Planning’ policy (section ‘General Sign Provisions’). We retained the typical visual look and feel of traditional chalkboards by printing a greenish washed-out background pattern. This material also proved to be much easier to install, was cost-effective, weather-resistant, and is certified as being completely recyclable.

The aim of our design was to augment the existing terrace house architecture with a striking visual accent that was physically adapted to the aesthetics of the existing environment. We took inspiration from similar boards that are commonly used for real estate notices or commercial advertising, while avoiding the visual occlusion of cars or trees. Our display was specifically designed to be able to fit to any façade width without affecting the overall layout: each display can be sized to perfectly fit with any existing balustrade, and it can fit the narrowest townhouse (< 3m), while still looking intentional and well-balanced for much wider houses (3–4.5m). We chose to print a permanent background template (including a grid and text) to provide for a strong recognizable visual framework on which the more messy handwritten information would visually stand out. These predefined graphic elements also made the display look more professional and less random, as the displays needed to convey a high level of trust to all stakeholders. The fixed background template also ensured a more

efficient updating process, as fewer elements had to be manually redrawn. In short, it was important to create a visual design that balanced a level of trust and seriousness with a degree of fun and happiness: a chalkboard on a residential façade – as simple as the physical components seem to be – is still an unusual combination and would indeed become a small neighborhood attraction.

3.2 Persuasive Visualization

The design of the data visualizations faced the complex dilemma of representing the differences in energy usage between individual households in a fair and honest way. Energy usage depends on many factors, such as the number of people living in the household, the inhabitable surface area, the type of appliances used, whether gas or electricity is used for heating, and so on. Any fair comparative feedback should ideally avoid showing factual usage data, and instead be based on some form of normalized data, i.e. the actual *change* in energy consumption over a specific period of time [29]. Focusing on change instead of actual electricity usage also provides for a certain degree of public privacy: communicating “*no, or little change*” conveys a different meaning from “*no, or little usage*”, for instance in the case when a household is absent for a long period of time. However, focusing on relative change does not take into account whether the usage is generally low or high in comparison to others, while it also hides any longer-term trend supporting the change. ‘Change’ also implies a comparison to a point in time, which is relatively difficult to determine. For instance, while comparing one’s energy usage to that of last week seems the fairest method (e.g. accounting for different living patterns in weekends versus weekdays), we deemed it difficult to invoke immediate comparison or encouragement (i.e. who actually remembers their activities of last week?). To ensure people were able to relate their recent decisions and activities to the actual performance shown on the displays, we preferred to compare daily (vs. weekly) energy usage: even when, for instance, most common households would show a ‘negative’ trend from Friday to Saturday, this trend would still be shared by most others and thus be comparable in a truthful way. In order to provide a more detailed view of the energy usage patterns in multiple varying positive, encouraging and visually attractive ways, we chose for the following combination of textual and graphical depictions (see Figure 2).

- a. **Marginal Notes** for personalized messages or persuasive captions, which could be written in the areas on the sides that were deliberately left blank.
- b. The **Daily Performance**, shown as a numerical percentage value, conveyed the change in energy consumption over the last 24 hours of a single household.
- c. The **Neighborhood Ranking** summarized the daily performance in terms of change of all participating households in a single numerical ranking. It aimed to encourage competition, while also reduced the need to compare individual displays.
- d. The **Historical Graph** showed the actual usage trends over time, and conveyed the time-varying, complex character of household energy consumption. The line graph was horizontally divided by weekdays, and was vertically normalized by the maximum usage measurement, mainly for privacy reasons. The graph was never wiped out, but extended at each update. The data for successive weeks

- were drawn on top of each other in different colors, so that the historical performance of the same household could be easily compared and contrasted.
- e. The **Pictorial Bar** highlighted the occurrence of any sustained change, such as a succession of positive or negative changes in energy use by way of an explicit visual reward. The pictorial system acted similarly to common stock market iconography: it only counted ‘better’ or ‘worse’ performances based upon pre-defined thresholds (e.g. it ignored changes less than 10%), but still recognized sustained lower energy usage levels, even after drastic changes. A reward or discouragement was depicted as a simple emoticon, which conveyed a degree of negative or positive emotions. Each day of the month, an emoticon was added within a preprinted circle. Facial emoticon expressions were chosen, as they are universal and intuitive to comprehend.
 - f. **Private Display**. Households were also provided with a common electricity monitor, embedded in a custom-made blackboard (Figure 5, Right). We decided on giving explicit access to the monitor device in order to entice the trust that the information shown on the public display was based on continuous and accurate measurements. The small, custom-made blackboard aimed to encourage participants to take notes about appliance usage or other energy-related observations, similar to keeping a journal, which can support self-reflection and in turn, the discovery of intrinsic factors of motivation [24]. The private display was pre-printed with weekday abbreviations to accentuate the daily update cycles embodied by the public display.



Fig. 2. Overview of the graphical depictions on the public feedback display: Marginal Notes (a), Daily Performances (b), Neighborhood Ranking (c), Historical Graph (d), Pictorial Bar (e).

3.3 Technical Implementation

Our extensive design efforts resulted in a fairly simple technical implementation. A commercially available Efergy E2 device, consisting of a sensor, wireless transmitter and wireless monitor, was used for measuring the actual electricity consumption. The monitor captures the energy consumption per hour, which can be downloaded via a USB connection. To ensure easy and uninterrupted access to this data, we left one dedicated monitor outside the house, e.g. in the electricity box or in a rainproof plastic

container (Figure 3, Left). To download the electricity data, we used a Netbook running Windows 7 (Figure 3, Center). The public feedback displays were produced using Corflute®, a light material made of twin wall polypropylene sheets (with an average size of 3.3x0.8m). Simple eyelets were used to mount the sheets onto the terrace fences with standard cable ties. The content was written and drawn using standard liquid chalk pens, which have the advantage over common chalk to be fairly water resistant. A 14-tread ladder was necessary to reach the displays (Figure 3, Right). A second Efergy E2 wireless monitor was used for the private display. The blackboard that encompassed the private display unit was custom-made with a size of 26x23cm and a chalk tray attached to the front (Figure 5, Right). The blackboard came with a detachable stand and a small hook for placing or hanging the display anywhere inside the home. The display could be set to reveal instant kilowatts, carbon emissions or costs per hour (with an update rate of 8s), as well as show the historical trends for the previous seven days.



Fig. 3. Left: Hidden storage of a separate electricity monitor allowing uninterrupted access to the data. Center: Netbook and electricity monitor used for downloading and updating displays. Right: The manual update of the daily neighborhood ranking using a ladder.

4 Methodology

4.1 Setup

For our study, we aimed to recruit about twelve households, six for the treatment (with public display), and two control groups of three households each (see Section 4.2). First, we selected two streets in a high pedestrian traffic area, featuring terrace houses of comparable size. Both streets were located in the vicinity of our university, in order to accomplish the manual updates more efficiently. Some households were selected based on the fact they were located on an intersection, so that the house façades were visible from a longer distance (Figure 1, Right). We used leafleting and door knocking to recruit households [35], which also allowed us to experience the study environment as a rich physical and social context, observe the actual community, and learn about the local attitudes regarding the topic of sustainability. The Efergy E2 electricity monitor (valued AUD\$134) was offered as an incentive to participate. We then physically visited the neighborhood at times when it was most likely that people would be at home, and systematically knocked on all houses that featured

a terrace and were of comparable size, and had no trees or other obstacles blocking the view. After returning four times, we recruited six households in each street, and arranged one dedicated contact person per household. Just before commencement, two households dropped out for individual reasons, one of which was replaced, leaving us with a total of 11 households (see Table 1).

Table 1. Participating households, their characteristics, and conditions (Group A: public and personal displays; Group B: private display; Group C: no feedback)

No	Occupants	Ownership	Group	No	Occupants	Ownership	Group
H1	2 adults (couple)	Owners	A	H6	2 adults, 1 child	Owners	B
H2	2 adults (couple)	Owners	A	H7	2 adults (couple)	Owners	B
H3	3+ adults (students)	Renters	A	H8	2 adults (couple)	Owners	B
H4	3 adults (shared)	Renters	A	H9	6 adults (students)	Renters	C
H5	3 adults (shared)	Renters	A	H10	2 adults (shared)	Owners	C
				H11	3+ adults (students)	Renters	C

4.2 Evaluation Study

We used a between-subject design study with three separate conditions. Two of the conditions were used as control measure. In condition A (n=5), both the public and private displays were installed. In condition B (n=3), only the private display was installed, which offered the same features as the private display used in condition A, but without the blackboard for note-taking. Condition C (n=3) was identical to B, but the energy monitor was not made accessible or visible to the participants. The continuous energy measurements ran for a total period of 10 weeks, during which the public displays were updated for 7 weeks. We interviewed all participants at the beginning and at the end of the study.

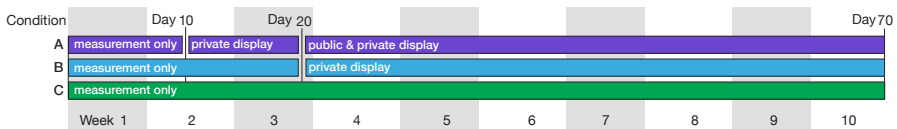


Fig. 4. Duration of measurement phases for the three between-subject study conditions.

Pre-study Interviews. Each pre-study interview (taking between 6-23 minutes) was conducted with the contact person from each household, in their house. Where possible we made arrangements that all other household members were present as well. In the interviews, we collected general information about the household, attitudes towards global warming, privacy, and the relationship with neighbors.

Energy Measurements. The study used a total of 16 electricity monitors: one per household in conditions B and C, and two per household in condition A (i.e. one inside and one hidden outside). As shown in Figure 4, in condition A, the energy data collection started for a period of 10 days, after which the private displays were installed. After another 11 days, we installed the public displays and continued

recording for 7 weeks. In condition B, we measured without intervention for 21 days, after which the private displays were made available for the remaining 7 weeks. In condition C, we measured without intervention for the entire period of 10 weeks.

The energy usage data was cleaned to eliminate obvious errors or outliers, for instance for two separate cases where the electricity monitor experienced a technical failure (see Table 2). Any erroneous (i.e. 0.0kWh) or outlier (i.e. a household consumption of less than 75% of the average consumption of the same weekday) measurement was replaced with the average value of correct measurements for the same weekday over the entire respective study phase. Outliers represented abnormal situations, such as participants being absent for one or more days, or a significant change in the number of people staying in the house. This systematic approach for cleaning the data was tested and validated against the anecdotal evidence from households where we were aware of the exact dates of abnormal events.

Table 2. Errors and outliers per household in number of affected days

Household	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
Errors	-	-	-	11	-	-	-	10	-	37	-
Outliers	19	15	11	14	9	9	7	8	11	2	10

Display Updates. The public displays were updated every single day for a period of four weeks, after which the historical graph area was wiped clean to avoid visual clutter. The graph from the fourth week was then reapplied before continuing with the updates. We used a custom-made software tool that visually simulated the public display, to determine the exact graphical depictions based on the actual sensor data. In total, each manual update session took between 1.25 and 2 hours (or 15 to 24 minutes on average, per house). The update time depended on the number of researchers involved (varying between one and two), the day of the week (with updates in the beginning of the week leading to increased repositioning of the ladder), and weather conditions. The updates generally took place in the late afternoon. The public displays were not updated during four distinct days, due to heavy rain conditions.

Observations. While we were present in the neighborhood, we kept a journal of observations and conversations with any participant or passer-by. We occasionally asked passers-by who approached us some informal questions regarding their understanding and perception of the displays.

Post-Study Interviews. After the study, we conducted another round of semi-structured interviews with all participants in their respective homes. The questions covered their awareness of energy consumption behavior and their perception of the displays. Participants also rated the public and private displays regarding their attractiveness, usefulness, ambientness, and enjoyment. These interviews took approximately one hour for condition A, and 6-12 minutes for conditions B and C.

5 Results and Discussion

5.1 Awareness

During the pre-study interviews, all households stated they were well aware of the climate crisis. Everyone expressed some opinion on global warming and most participants stated that they were trying “*to do their bit*”, which included switching off lights (n=11), switching off appliances at the power plug (n=6), replacing bulbs with energy-efficient lighting (n=3), replacing electric water heating or stoves with gas ones (n=2), and partly switching to green energy (n=2). None of the households was regularly checking their electricity meter to monitor consumption. However, all households were monitoring their energy bills regularly, at least to check the total costs, while some also checked other data provided on the bill, such as greenhouse emissions. Nine households stated that they compared their consumption from bill to bill, mainly to see whether the costs had increased, and then would decide whether to look into the information provided on the bill in more detail. Only one household (H10) had actually compared their consumption with those of others.



Fig. 5. Left: Neighboring public feedback displays during the last week of the study. Right: A private feedback display embedded in blackboard showing notes on appliance usage.

In the post-study interviews, we identified several anecdotal reports about the effect of the feedback displays on the general awareness of energy consumption. One household was using the private display blackboard to take handwritten notes about appliance usage (i.e. of kettle and laptop, and even of the difference between a hot and cold washing cycle), while nobody used the pre-populated grid of weekdays. Another household invented a game using the blackboard during which the goal was to turn on or off appliances in order to reach a certain score, which they then recorded on the blackboard. This household described the private display as “*a point of conversation for us, bring[ing] us together*”, to which was added: “*bringing us together as a team*”, referring to the neighborhood ranking, where they saw themselves competing as team against the other households.

While the private display affected the awareness of energy consumption on an appliance level, we found that the public display allowed participants to understand their consumption on a more general level. For instance, several participants noticed how

the patterns on the historical graph related to their behavior: “*The graph was interesting, because you could see that we had patterns ... you know, always take showers at the same time and we did see when we took a shower because of the hot water thing*” (H1); or noticing spikes: “*We don’t have routine anyway, we never do the washing until we run out of underwear and then we do four loads in one go*” followed by the participant’s partner’s comment: “*Then the following day on the graph, we see that it goes BEEP – ‘oh that’s our washing day’*” (H2). While the historical graph was mentioned as being most useful for identifying patterns, numerical indications also played an important role. For instance, one household (H5) mentioned how they attributed a 70% increase to the fact that they did three wash loads that day. Another household (H2) related a sudden negative neighborhood ranking to a similar event. Most participants reported how the public display prompted many conversations on energy consumption and environmental issues with other household members or people visiting and noticing the public display.

5.2 Behavior Change

During the post-study interviews, households generally stated that they perceived the private display as more influential to their behavioral changes than the public display, mostly because of its real-time nature. These findings (in conditions A and B) are in line with other studies on smart energy monitors, which found that all participants reported changes [14]. In particular, the monitoring device led to discoveries relating to the electricity consumption of individual appliances. As such insights have been previously reported, this section instead focuses on the impact of the public display.

We observed that the public display induced or reinforced behavior change. For instance, one household mentioned how the patterns observed from the private display correlated with those on the public display, which led to the suggestion to replace the water heater and a form of long-term and sustained conservation behavior: “*I would try to use less and less and less [to come first], so in the beginning we were often number one; ... but then I realized that I was just using less and less every day, which is not going to work out, because eventually I needed to do the laundry*” (H1). While this household became frustrated by the ranking system being based on change rather than factual consumption, they still showed the highest conservation behavior by consuming 25% less energy during the first week, maintaining this conservation for another week, after which they gave up and increased their consumption by 21%.

Behavior change was often triggered by the competitive nature of the public displays, and in particular by the neighborhood ranking. However, the neighborhood ranking also led to less sustained behavior, like strategically clustering washing cycles in time, in order to end up first in the ranking the day after (e.g. H2). In another case, the neighborhood ranking triggered more spontaneous short-time behavior change: “*One day we were [away] and got a message from our housemate saying ‘we are number five..., the people next to us are number one!’ and I sent back a text [telling them] to quickly switch of all our power points!*” (H5). We also observed the excitement about the competitive aspect during our manual update sessions, when we often noticed occupants coming out of their house after we finished updating their display to check their own ranking. For instance, we once noticed how H4 checked

their ranking (first that day), and then ran back into the house announcing their achievement loudly and proudly to the other inhabitants.

Overall, findings from the interviews suggest that the private feedback led to valuable intrinsic insights, which allowed participants to develop long-term strategies for reducing their consumption. In contrast, the more competitive aspect of the public display was seen as an extrinsic factor of motivation, as participants often aimed to reduce their consumption in order to ‘win’: one household described how they specifically focused on the ranking, since they “*wanted to come first*” (H2), while H4 revealed that they paid most attention to the neighborhood ranking. Overall, the ranking aspect worked well to encourage people to ‘start’ with changing their behavior, although the behavior seemed to decline after people got used to its competitive aspect. One household commented on how there was definitely more behavior change noticeable due to the neighborhood ranking after the public display went up, but that at the end they tended to pay less attention to the ranking since it changed almost daily (H5). This finding suggests that the design of the pictorial bar seemed to have failed in encouraging longer-term or sustained behavior. A possible solution could consist of adding ranking variations, such as adding weekly or monthly charts, so that long-term goals would be more explicitly and intuitively recognized than the graphical nature of the emoticons. Only one household mentioned that they did not pay much attention to the neighborhood ranking since they found it less useful, yet they considered the ranking to be “*cool*” and actually did look at the displays of other households. One should note competitive behavior is not for everyone, an important sentiment any comparative feedback should be aware of.

5.3 Energy Conservation

A data analysis of weekly consumption per condition revealed patterns related to the feedback. Figure 6 shows a considerable reduction of energy usage in week 4, around which households received their feedback displays (Condition A: -13.2%; Condition B: -12.1%). Households in condition A maintained their conservation behavior for the following week (-0.4%), while condition B showed an increase of energy usage (+10.8%). This might be best explained by the decline of the electricity monitor’s novelty factor, or because many naïve conservation strategies are difficult to sustain longer-term (e.g. washing clothes less often). These reasons might also explain the energy usage increase of condition A in week 6 (+5.7%). In this week, household H1 stands out: they came to understand how the ranking was based on change, which they described as “*competing against yourself*” and gave up competing. The consumption of both control groups (conditions B and C) remained approximately the same throughout week 6 (-3% and -1.6%) and week 7 (-1.4% and 0.7%), while the condition A’s consumption was declining in week 7 (-7.3%). For the remaining weeks, condition A’s consumption remained almost unchanged or even declined further (between -0.2 and -1.1%). The little change in these three weeks could be attributed to the data corrections in week 9 and 10 for H1 and H4, which accommodate some technical failures with the measurements. However, an analysis of the data excluding H1 and H4 showed little effect with the overall decrease in week 8-10 changing -3% or larger. Condition B showed a very similar usage pattern, with an increase in week 7 followed by two weeks of considerable to little decrease. It is

unclear what caused the increase in week 8 in condition C, which was however counterbalanced by a constant decrease over the following two weeks. The remarkable decline in energy usage for conditions A and B in week 2 could possibly be attributed to an increase of the minimum ambient temperature during that week, as compared to the previous week. However, as most households from condition A stated that they would rarely to never use electric heaters, this change is best explained by receiving the private display in that week. It is unclear what caused the increase in week 3, and while it might have naturally led to a decrease in week 4, condition A maintained this level throughout week 5, so that it can be attributed to the public display. The average change per week over the deployment period (week 4-10) was -2.5% for condition A, -1.0% for condition B, and -0.5% for condition C.

The energy consumption before and after the installation of the feedback displays shows a decrease in usage from week 3 to week 4 of -13.2% for condition A and -12.1% for condition B, while condition C remained the same. Interestingly, the conservation performance of condition B declined towards the end of the study period with an increase in energy consumption of $+8.5\%$ in week 10 compared to week 3, while condition A decreased their usage further to the amount of -17.0% (condition C: -4.3%). This phenomenon implies that households with public display positively changed their energy consumption behavior, while the traditional real-time energy monitor worked for short-term change only. However, due to the small sample size, these results can only be considered to show trends. A one-way ANOVA test did not reveal any significant differences, which can be attributed to the small sample size.

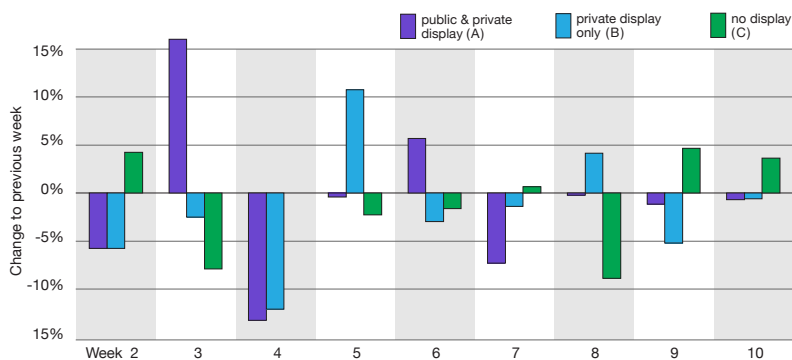


Fig. 6. Weekly change in energy consumption per condition (corrected measurements)

5.4 Representation Methods

Only one household (H5) used the small blackboard that came with the private display for its intended purpose, yet all households wanted to keep the blackboard. H2 even removed the monitor from the blackboard to place it closer to their appliances in the kitchen, while they placed the blackboard near the phone for general note taking. Others stated they might eventually start using it together with the electricity monitor after the study. Not everyone appreciated the aesthetics of the private display, with H4 stating that they thought it looked “tacky” and “ugly”, and the public display was

“attractive” and “playful” compared to it. In general, we received varying responses regarding the design of the displays, with some households judging the private display to be more aesthetic, while others rated the public display higher. H2 and H5 particularly mentioned that they liked the visual style of the chalk colors, H2 added how they appreciated the way the display became messier and messier over time. H5 described how they thought that the display made their house look “*more modern*”. Similarly households’ responses regarding the ambientness of the displays varied. H4 stated that “*you got used to it being on the house and then it’s just there*”. Others rated the ambientness low, saying that the display was not designed to be unobtrusive.

The different visualization techniques on the public display were rated positively. However, H2 found the emoticons on the pictorial bar “*too judgmental*”: they even wiped out the neutral face they received on the first day and left a message for us saying “*no more sad faces please*”. We consequently changed their pictorial representation to flowers, with different colors and sizes representing different levels of energy conservation performance, similarly to the original emoticons. When asked in the final interviews, none of the other households expressed any concerns regarding the emoticons. Some households found the fact that the display showed change rather than factual usage counter-intuitive. All but one stated that they would have preferred factual data to check how their actual energy consumption compares to their neighbors’. These attitudes went straight against our initial assumptions, as nobody seemed concerned about the severe privacy implications of displaying energy usage data, which might be explained by the playful nature of the study. Basing the ranking on change also led to some confusion regarding the neighborhood ranking, so that as H1 was initially puzzled about being ranked third after being away for three days. Although they came first in the ranking the day after they left, there was no change in consumption for the following 2 days of their absence, which thus resulted in a low overall ranking. While the pictorial bar was meant to solve such issues by visually rewarding persistence and sustained good behavior, its understandability suffered from the fact that all other elements were updated in the late afternoon for the previous 24 hours, while the emoticon represented a summary of the entire day.

6 Conclusion

Our quantitative measurements indicate the apparent potential of public visualization and neighborhood competition to encourage behavioral change. Households that received a public display for a period of 7 weeks decreased their energy usage on average by 2.5% per week (compared to a decrease of 1.0% in control group B and 0.5% in control group C). The presence of a public display further led to a more sustained conservation behavior compared to only having access to private feedback. Our qualitative results confirmed the effect of the competitive neighborhood ranking as being ideal for initiating behavior change. However, the competitive aspect also led to several unexpected side effects (e.g. clustering energy-intensive activities to specific strategic times). All stakeholders appreciated the competitive features, and the expressive playfulness of the visual design. Unfortunately, it became clear that the single, numerical ranking played a too overwhelming role in motivating people into short-term competitive behavior, while the pictorial bar, a feature that was specifically

designed to support sustained behavior, was not sufficiently accepted as a longer-term reward. Although the historical graph could be exploited to reveal daily activities such as showering and washing, there was little to no notion of privacy concerns expressed by the participants. The chalkboard-like approach proved to be successful in keeping deployment costs low, but also allowed us for ad-hoc personalization (such as replacing emoticons with flowers when people disliked them, or adding written explanations when we noted how some households misunderstood the data). The fact that the public display was updated only once a day was seen as less a problem than we expected, which could be attributed to how the private displays compensated for more 'immediate' information needs. The manual updates turned out to be an important aspect of the study, since they required us to visit the site daily, allowing us to come into contact with households and passers-by. These informal conversations provided a rich source of feedback regarding the performance of our system and helped us to better understand the driving principles behind the data.

However, we greatly underestimated the required effort to update the displays on a daily basis, as it took one to two researchers more than an hour per day to complete the task of updating only 5 house facades. The update process was complicated by the physical height as well as the relatively large amount of visual elements that had to be drawn. At the start of the study, households were often confused regarding the meaning of certain visual elements. Passers-by – while recognizing that the display was about energy consumption – often asked questions about the meaning of some depictions. Providing explicit explanations to the public about the visualizations could possibly have helped the overall performance of the public display. In this study, we approached public feedback from a more realistic scenario of a council or energy supplier installing publicly accessible information boards on houses. However, even when we used customizable chalkboards, a more participatory design process involving the community might have led to a more successful design that invoked stronger feelings of ownership. Based on the generalized findings from this study we suggest the following design considerations for public representation of energy usage:

Privacy. Even though our study did not discover any apparent privacy concerns, the visualization of energy usage data inevitably reveals when inhabitants are at home or not, when they shower, or when they go to sleep. Possible solutions are: 1) removing detail: for instance by averaging the data over longer time periods; or 2) normalizing the usage data, which still reveals some detail but removes the possibility to compare the size of trends from one occurrence to another, or from one household to another. The biggest issue, however, is that people will show a very strong tendency to assume any time-based graph to represent real usage, even when the data is averaged or normalized, and labels, legends or captions provide other information.

Fair. A visualization should compare variables that are indeed comparable, taking into account influential differences (e.g. size of household, inhabitable area). Therefore, showing relative changes instead of real measurements seems most suitable, but they are difficult to intuitively understand. For instance, people need to relate to a historical time period that is never really representative or objectively comparable (e.g. even contrasting weekdays instead of successive days suffers from differences in weather). Numerical rankings might be a possible solution, as they can hide more

complex formulas from view. However, people cannot easily relate rankings to personal effort, as other people's performances come into play.

Trust should be conveyed, in that people believe that the data shown is indeed accurate. Trust can be typically gained by way of appearance (e.g. from expensive or professional materials), direct associations (e.g. associating a trustworthy organization with the display), or persistence (e.g. updates occur regularly and without fail).

Encourage and Sustain Change. While access to information provides insight and motivation in one's behavior, some form of visceral experience is required to support the sustainment of positive behavior over longer periods of time.

Ability. It is one thing to become more aware of one's behavior, and to be enticed to change, it is another to know what action to take. Clear indications should be provided of 'how' to change behavior, to avoid feedback to become frustrating.

Immediate. While behavioral change should ideally be made apparent as quickly as possible, people seem patient to be informed of their competitive performance when also having access to alternative means of immediate feedback.

Understandable. Many issues relate to conveying complex, time-varying data to lay people, including visual density (i.e. showing not too much or not too little information), cultural sensitivity (e.g. emoticon iconography might lead to different sentiments), and aesthetic preference (e.g. relating to the sensitivity of one's own house façade). Here, it might help to allow personalization in terms of aesthetics, or deliberately restrict the overall visual complexity to basic elements.

Collaboration and Competition. A public display on behavior should be sufficiently dynamic and allow different forms of competitive behavior to ensure continuous and persistent interest. Due to the long-term goals, public displays of energy consumption need to incorporate elements of both short-term and long-term comparison, in order to facilitate immediate as well as sustained saving, while minimizing the long-term effects of exceptional circumstances (e.g. households being away, starting competing later). However, any public competition should emphasize a playful and positive character to induce cooperation and "friendly" discussion, rather than envy and shame, by establishing a common ground amongst all stakeholders.

Our study indicated the potential as well as the many challenges for the public visualization of household energy usage. It adds new knowledge to the field of urban computing, particularly regarding the real-world use of communal and sociable screens in the urban domain. In the future, we foresee similar applications, either more minimalistic and integrated in the façade typology, or more personalized and free to accommodate more individual approaches towards sustainable living.

Acknowledgments. We would like to thank all participants in the study; the City of Sydney for their support; Josh McInerheney and Damien Kwan for helping with the updates; and Dan Hill for his advice. This research was funded by the Sustainability cluster at the Faculty of Architecture, Design & Planning of the University of Sydney.

References

1. International Energy Outlook 2007. Office of Integrated Analysis and Forecasting, US Department of Energy, Washington, DC (2007)
2. Breslau, K.: The Resurrection of Al Gore. *Wired* 14 (2006)
3. Egan, C.: Graphical Displays and Comparative Energy Information: What Do People Understand and Prefer? In: Summer Study of the European Council for an Energy Efficient Economy (1999)
4. Holmes, T.: Eco-Visualization: Combining Art and Technology to Reduce Energy Consumption. In: *Creativity and Cognition*, pp. 153–162. ACM, New York (2006)
5. Roberts, S., Humphries, H., Hyldon, V.: Consumer Preferences for Improving Energy Consumption Feedback. Report to Ofgem, Centre for Sustainable Energy (2004)
6. Pierce, J., Schiano, D.J., Paulos, E.: Home, habits, and energy: examining domestic interactions and energy consumption. In: *Proc. CHI 2010*, pp. 1985–1994. ACM, New York (2010)
7. Darby, S.: The Effectiveness of Feedback on Energy Consumption. *Env. Change Inst.*, Oxford (2006)
8. Staats, H., Leeuwen, E.v., Wit, A.: A Longitudinal Study of Informational Interventions to Save Energy in an Office Building. *J. Applied Behav. Analysis* 33(1), 101–104 (2000)
9. Fogg, B.J.: *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, San Francisco (2002)
10. Froehlich, J., Findlater, L., Landay, J.: The design of eco-feedback technology. In: *Proc. CHI 2010*, pp. 1999–2008. ACM, New York (2010)
11. Consolvo, S., McDonald, D.W., Landay, J.A.: Theory-driven design strategies for technologies that support behavior change in everyday life. In: *Proc. CHI 2009*, pp. 405–414. ACM, New York (2009)
12. Fogg, B.J.: A behavior model for persuasive design. In: *Proc. Persuasive 2009*. ACM, New York (2009)
13. Mountain, D.: *The Impact of Real-Time Feedback on Residential Electricity Consumption: The Hydro One Project* (2006)
14. Hargreaves, T., Nyea, M., Burgessa, J.: Making energy visible: A qualitative field study of how householders interact with feedback from smart energy monitors. *Energy Policy* 38(10), 6111–6119 (2010)
15. Ueno, T., Inadab, R., Saekib, O., Tsjib, K.: Effectiveness of an Energy-Consumption Information System. *Applied Energy* 83(8), 868–883 (2006)
16. Wood, G., Newborougha, M.: Energy-use information transfer for intelligent homes: Enabling energy conservation with central and local displays. *Energy and Buildings* 39(4), 495–503 (2006)
17. Martinez, M.S., Geltz, C.R.: Utilizing a Pre-Attentive Technology for Modifying Customer Energy Usage. Paper presented at the European Council for an Energy Efficient Economy (2005)
18. Ham, J., Midden, C.: Ambient Persuasive Technology Needs Little Cognitive Effort: The Differential Effects of Cognitive Load on Lighting Feedback Versus Factual Feedback. In: Ploug, T., Hasle, P., Oinas-Kukkonen, H. (eds.) *PERSUASIVE 2010*. LNCS, vol. 6137, pp. 132–142. Springer, Heidelberg (2010)
19. Jacobs, M., Löfgren, U.: Promoting Energy Awareness through Interventions in Public Space. In: *Proc. NordCHI 2005*. ACM Press, New York (2005)
20. Gyllensward, M., Gustafsson, A.: The Power-Aware Cord: Energy Awareness through Ambient Information Display. In: *EA CHI 2005*. ACM, New York (2005)

21. Pierce, J., Odom, W., Blevis, E.: Energy Aware Dwelling: A Critical Survey of Interaction Design for Eco-Visualizations. In: Proc. OZCHI 2008. ACM, New York (2008)
22. Vande Moere, A.: Towards Designing Persuasive Ambient Visualization. In: Issues in the Design & Evaluation of Ambient Information Systems Workshop, pp. 48–52 (2007)
23. Björkskog, C., Jacucci, G., Mikkola, T., Bertoncini, M., Gamberini, I., Torstensson, C., Nieminen, T., Briguglio, I., Andriani, P., Fiorentino, G.: BeAware: A Framework for Residential Services on Energy Awareness. In: Proc. UBICOMM 2010 (2010)
24. He, H.A., Greenberg, S.: Motivating Sustainable Energy Consumption in the Home. In: Defining the Role of HCI in the Challenges of Sustainability Workshop (2009)
25. De Young, R.: Changing behavior and making it stick: the conceptualization and management of conservation behavior. *Env. & Behavior* 25(3), 485–505 (1993)
26. Abrahamse, W., Steg, L., Vlek, C., Rothengatter, T.: A review of intervention studies aimed at household energy conservation. *J. Env. Psyc.* 25, 273–291 (2005)
27. Brandon, G., Lewis, A.: Reducing household energy consumption: A qualitative and quantitative field study. *J. Env. Psyc.* 19(1), 75–85 (1999)
28. Foster, D., Lawson, S., Blythe, M., Cairns, P.: Wattsup?: Motivating reductions in domestic energy consumption using social networks. In: NordiCHI 2010. ACM Press, New York (2010)
29. Petkov, P., Köbler, F., Foth, M., Krcmar, H.: Motivating domestic energy conservation through comparative, community-based feedback in mobile and social media. In: Proc. C&T 2011. ACM, New York (2011)
30. Make Me Sustainable, <http://makemesustainable.com/>
31. Carbonrally - Green Living, <http://www.carbonrally.com/>
32. Google Powermeter, <http://www.google.com/powermeter/>
33. Kaufman, L.: Utilities Turn Their Customers Green, With Envy. *The New York Times* (2009), <http://www.nytimes.com/2009/01/31/science/earth/31compete.html>
34. Fatah gen. Schieck, A.: Animate Space: Urban Environments as Medium of Communication. In: Proc. Space Syntax Symposium (2005)
35. Davies, K.: Knocking on doors: recruitment and enrichment in a qualitative interview-based study. *Int. J. of Social Research Meth.* (2010)

Are First Impressions about Websites Only Related to Visual Appeal?

Eleftherios Papachristos and Nikolaos Avouris

Human-Computer Interaction Group, Electrical and Computer Eng. Dept.,
University of Patras, GR-265 00 Rio Patras, Greece
{epap, avouris}@ece.upatras.gr

Abstract. This paper investigates whether immediate impression about websites influences only perceptions of attractiveness. The evaluative constructs of perceived usability, credibility and novelty were investigated alongside visual appeal in an experimental setting in which users evaluated 20 website screenshots in two phases. The websites were rated by the participants after viewing time of 500 ms in the first phase and with no time limit in the second. Within-website and within-rater consistency were examined in order to determine whether extremely short time period are enough to quickly form stable opinions about high level evaluative constructs besides visual appeal. We confirmed that quick and stable visual appeal judgments were made without the need of elaborate investigations and found evidence that this is also true for novelty. Usability and credibility judgments were found less consistent but nonetheless noteworthy.

Keywords: Webpage design, aesthetic evaluation, credibility, visual appeal, perceived usability.

1 Introduction

The importance of appropriate aesthetic web design has been clearly shown by Lindgaard et al. [1] in a series of experiments about the immediacy of first impressions. Their findings indicate that first impressions about websites can be formed during the initial 50 ms of viewing and that they are highly stable over time. In a subsequent study Tractinsky et al. [2] replicated and extended the above study providing further evidence for the immediacy and consistency of aesthetic impressions.

These results had quite an impact on the HCI community because they suggested an elevated importance for website aesthetics. However, there is an ongoing debate about the nature of such aesthetic responses regarding the involvement of cognition. According to Norman [3] the visceral response to visual stimuli is merely an affective unconscious reaction about good or bad: a “gut” feeling. Hassenzahl [4] rejects the notion of visceral beauty stating that beauty judgments are “cognitive elaborations of the initial diffuse reaction” to stimuli. In that vein of thought cognition is required for aesthetic judgment. Additionally, that initial reaction may serve as a starting point for subsequent, more complex evaluation which often involves expectation and prior

experience. However, Lindgaard's et al. [1] and Tractinsky's et al. [2] results contradict to some extent the above by showing that their subjects could provide stable aesthetic evaluations in time periods too short to discern all of the stimuli details.

If first impressions are only positive or negative feelings about stimuli as Norman [3] and Hassenzahl [4] presume, then users wouldn't be able to distinguish between a set of high-level evaluative constructs. Any judgment would be a "halo effect" or a carry-over effect of that positive or negative impression to the other construct and evaluations should be highly correlated and not independent. If however, website users have predisposed concepts such as simplicity, symmetry and familiarity associated for example to usability perceptions then it is possible that judgments are a result of those individual intuitive criteria. If that is true then first impressions are not simple assessments of positive or negative feelings toward stimuli, but a bundle of quick and intuitive evaluations of several characteristics which are particularly important to the individual user.

However, from a designer's point of view it is important to understand the implications of website users' first impressions regardless of the origins of their formation. Are first impressions only about visual appeal? And if not, what else are users able to form opinions about in split seconds? In order to investigate if website users are able to form stable judgment about several website characteristics in a glimpse of an eye we had first to identify evaluative constructs previously linked to aesthetic matters. Literature research helped us identify: perceived usability [5,6], perceived credibility (trustworthiness) [7] and novelty [8,9] as appropriate constructs for the purposes of our study.

The objectives of this study were:

1. To investigate whether the formation of impressions about other high level evaluative constructs related to aesthetic design (perceived usability, credibility and novelty) is as quick as visual appeal, and how stable they are over time.
2. To examine whether their judgments on the evaluation constructs for the websites are independent or only covariations with visual appeal.

2 Method

Forty undergraduate university students (25 male, 15 female, aged 21 – 34, mean age = 23.9) participated in the study as partial fulfillment of the requirements in a human computer interaction course. The participants evaluated screenshots of 20 hotel websites. All participants reported having previous experience with hotel websites in general but none with the specific sample selected for the study. The selected hotel websites originated from a remote to the participants destination country (New Zealand) in order to minimize the possibility of prior sample familiarity. The website selection criteria were to have a balanced sample of good, average and bad designed websites. Although the selection process was subjective, post evaluation analysis showed that participants perceived our sample as balanced according to visual appeal. Unlike the studies of Lindgaard et al. [1] and Tractinsky et al. [2] we felt that our test material should belong to the same website domain in order to minimize possible confounding factors.

In addition, we had to reduce stimuli number to avoid participants' fatigue since they were asked more questions per website. Similar to Tractinsky et al. [2] we chose to replicate only the 500 ms condition of Lindgaard's et al. [1] experiment, which has been characterized as a time period short enough to form first impressions, but not long enough to evaluate other features such as semantic content [1][2].

2.1 Procedure

After participants were informed about the purpose of the experiment, specific instructions were given about the evaluative constructs (visual appeal, perceived usability, credibility and novelty) in order to ensure a unanimous understanding of them.

The evaluation took place on an eye-tracker (Tobii T60) using a specifically developed software. In the study's first phase the test websites were displayed as screenshots for 500 ms and were followed by a screen that contained the rating scales. We used an unmarked slider (from 0 to 100) as in [1] with the appropriate description on each end for each of the aforementioned evaluation criteria. Between each rating screen and each website screenshot a delay screen lasting for 1sec was placed. The delay screen contained a crosshair exactly in the middle of the screen in order to ensure that all users had the same viewing starting point. The software presented to each participant the website screenshots in a completely randomized order. There was no time limit while viewing the evaluation screen.

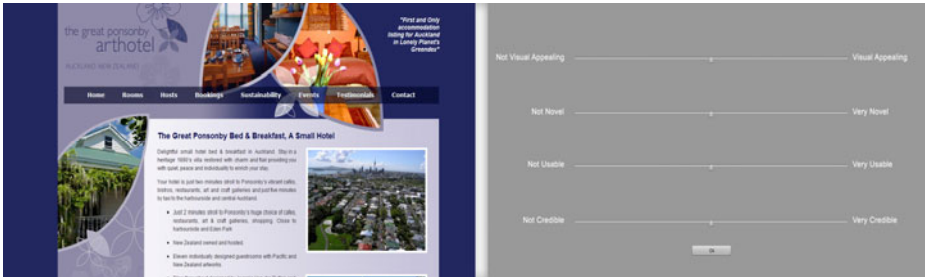


Fig. 1. Representative website (on the left), rating scales screen (on the right). A delay screen appeared between websites and ratings scales in each phase.

The second experimental phase was identical to the first one with the difference that there was no limit in displaying time. In this phase participants were asked to evaluate the same websites again on the same evaluation criteria after they viewed each website for as long as they wished. Screenshots were displayed in a new randomized order for each participant. The whole procedure lasted approximately 30 minutes for each participant.

3 Analysis

Since, 40 participants evaluated 20 webpages there was a total of 800 evaluations for each construct. As a first step of the analysis we examined the frequency distributions of

user evaluations for each construct individually. Our concern was to examine whether user evaluations revealed sample skewness in a particular construct which might limit result generalizability [2,6]. The examination showed quasi- normal distributions for all constructs, which means that most evaluations for the entire sample were around the middle of the scale and fewer at the extremes. In figure 2 mean rating's for the entire sample in both phases are displayed, all evaluations were more favorable in the second phase except visual appeal. However, visual appeal was the only construct with significant difference ($t(19) = 2.09, p = .05$) between the two phases.

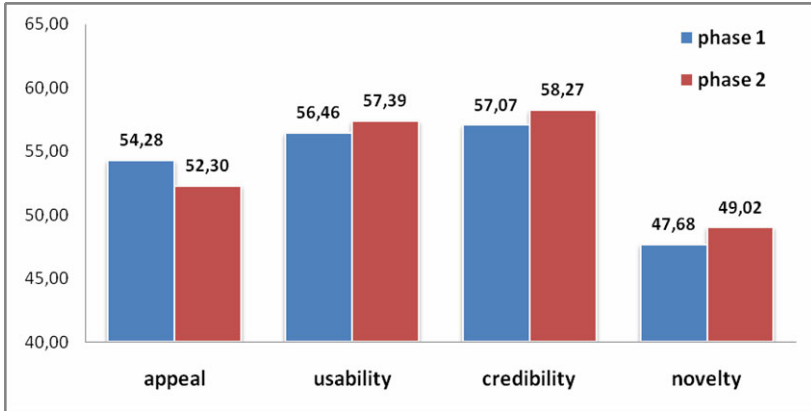


Fig. 2. Average evaluation of the entire sample in both phases

In a subsequent analysis we examined our data per website and looked at the correlations of average scores between the two phases for each construct. As it can be seen in table 1 the lowest correlations are for perceptions of usability ($r = 0.64$) with only 41.4% of the explained variance shared through the phases. The average ratings for the other constructs were highly correlated with explained variance ranging from 64.8% to 90.4%. Correlation of novelty perceptions were even higher ($r = 0.951, p < .001$) than of visual appeal. All correlations were significant and relatively high, indicating consistency for user evaluations between very short and long viewing periods when averaging over stimuli as in [1,2].

As a next step we looked into within-participants consistency by calculating the between phases correlations of each construct for each participant individually. In both [1] and [2] this analysis resulted in lower correlations than these aggregated over stimuli. Within-rater reliability, however, was notably lower in [2] (ranging from -0.09 to 0.9 with average correlation of 0.55) than in [1]. Our analysis yielded similar to [2] results indicating large variation in participant consistency (table 2). Participant reliability in visual appeal ratings ranged from $r = .03$ to $r = .90$ with an average of $r = .521$. The correlations of 13 participants fell below $r = .50$ and of 27 above. In total, 70% of the correlations were significant. Participant consistency was somewhat lower for the other constructs.

Table 1. Correlations of average scores

	Correlation	Sig.
Visual appeal	.864	.001
Per. Usability	.644	.002
Credibility	.805	.001
Novelty	.951	.001

Finally we investigated the between-construct relationship of websites mean ratings for each phase independently. In order for the constructs to have been judged independently we had to rule out that the evaluations were a result of simple covariation effects. If attractive websites were evaluated by participants high and the unattractive low on all constructs then the between construct correlation would be high. The ability of participants to differentiate between constructs, especially at the 500 ms condition, could indicate that evaluations are not simply a product of a positive or negative first impression. As depicted in table 3 credibility is positively correlated to visual appeal and to perceptions of usability in both experimental phases. The influence of visual appeal on perceived usability ($r = .45^*$) wears off in the second phase ($r = .19$) which could indicate that more elaborate investigation is needed by participants in order to form perceptions of usability.

Table 2. Within participants correlations

	Mean Correlation	Range	Sig. Cor.
V. Appeal	.521	0.03 - 0.90	70%
P. Usability	.261	0.01 - 0.92	22.5%
Credibility	.332	0.03 - 0.83	35%
Novelty	.503	0.10 - 0.90	62.5%

The interesting result, however, is that novelty is significantly negative correlated to perceived usability and to some extent to credibility, but is positively correlated to visual appeal (significant only in second phase $r = .49$). It seems that novelty perceptions, which were proven relatively consistent both in within participant and in average rating, mediate the other evaluations. As shown in [8,9] slightly above average novelty perceptions are associated with attractiveness, while extreme deviation from the norms results to confusion and therefore low perceived usability.

These results were a first indication that participants could differentiate at least novelty perception from a positive or negative first impression that was formed in split seconds. However average between – construct correlations alone is not enough to indicate independence of perception. For that reason we examined if particular websites received differing scores for all or some of the evaluative constructs. For example, finding some websites rated highly in perceived usability and at the same time low in visual appeal could indicate that the evaluative constructs were judged independently from each other.

Table 3. Correlations between constructs Phase A and B

Phase A	V. Appeal	P. Usability	Credibility	Novelty
V. Appeal	1	.448*	.580**	.284
P. Usability		1	.859**	-.608**
Credibility			1	-.395
Novelty				1
Phase B	V. Appeal	P. Usability	Credibility	Novelty
V. Appeal	1	.191	.484**	.489*
P. Usability		1	.713**	-.470*
Credibility			1	-.250
Novelty				1

For the 500ms condition aggregated over website ratings showed that the seven most appealing websites were also rated high in perceived usability and credibility but received only moderate novelty ratings. Six of the seven less appealing websites were rated low or average on novelty but high in perceived usability. Results from the second phase were very similar regarding the above trend, except from some minor changes in the ranking order of the websites. Although, on average none of the websites scored at the two extremes (largest difference was visual appeal=58.9 and perceived usability = 28.7) we found large divergence of certain constructs in individual ratings. Averaged over designs constructs scores were used as within subject’s variables in one way repeated measures ANOVA. The analysis revealed that construct differences were greater in the 500 ms ($F(1,19) = 8.82, p < .008$) than in the no time limit condition ($F(1,19) = 5.62, p < .028$). Post Hoc comparison showed that novelty ratings were significantly different from all other constructs in the first phase while only credibility and visual appeal differed significantly in the second. Although, constructs differences seem to vary between the two phases, the aforementioned results serve as first indicators of construct independence. However, further studies are required in order to fully understand their relationships.

4 Discussion

The above findings are indications that users form quickly reliable judgments about various websites characteristics. We found evidence that the formation of novelty perceptions in split seconds is particularly stable over time. It is certain that participants used inference and reflection while confronted with the rating scales since no time limit was imposed. Any kind of experimental setting can’t avoid tempering with the natural circumstances in which judgments about websites are made. Participants have to formulate their opinion or give ratings on a scale which interferes with the natural process in which websites are viewed, judged and used. However judgments made during extremely short and long exposure shared high explained variance which means that similar conclusions are made between having only glimpse and after

rigorous examination. In addition the judgments participants were able to make in this study are very different from simple reactions of liking or disliking.

On the other hand we found considerable differences in participants' ability to rate the websites under the experimental conditions. Three of them had no consistent rating in any of the evaluative constructs, most had only in one or two and only seven had significant correlations in all of them. The explanation for this could be that certain participants had strong, predefined notions about some constructs or strong likes or dislikes about design characteristics easily identified in the 500 ms condition (color, form, background texture). It is also possible that some participants had the ability to identify more visual attributes during the same timeframe than others.

The reliabilities concerning credibility and especially perceived usability were noticeably lower. Still the correlations reported aren't atypical in research in which human judgment process is involved. Although, we feel that usability judgments are more moderated by novelty we have to further investigate other alternative visual factors such as symmetry, complexity and order which have been previously linked with perceptions of usability.

In addition, we confirmed that average visual appeal evaluations of web pages are very consistent. Furthermore, within participant consistency was considerably lower than [1] and similar to [2]. An explanation for that could be that Lindgaard [1] used a polarized sample; half the websites were "ugly" and half "beautiful". In addition, in experimental phases 2 and 3 of their study, a subset of the initial sample was used after keeping only the websites that were rated on the extremes by users in phase 1. In our and Tractinsky's [2] studies the sample was indented to be balanced in terms of attractiveness-beauty by following a quasi - normal distribution. As previously indicated by several studies [10,11] and clearly demonstrated by Tractinsky et al. [2] in the same context (website evaluation) extreme ratings are more easily generated by participants. Apparently, participants need more time to evaluate close to average stimuli since more elaboration is needed to identify flaws or positive characteristics before forming a final opinion.

5 Conclusion

The present study was able to replicate findings of [1,2] regarding the consistency of visual appeal evaluations of websites between extremely short and long exposure. Our aim was to extent previous research and to investigate the consistency of additional evaluative constructs related to website aesthetics. We found indications that participants were able to provide stable ratings for novelty and to some extent for credibility and perceived usability. Our findings support the initial hypothesis that besides attractiveness other aesthetic responses are also able to be made by website user in the first critical split seconds of first viewing.

As future work we indent to analyze the eye-tracking data gathered during the experiment in order to examine what participants were able to focus on during the 500 ms period. Also, implicit measures such as response latencies for each evaluative construct could, as in [2], further validate our results. Finally, we intend to investigate the relation of low level constructs such as symmetry, order, complexity balance and contrast to the high level constructs investigated in this study. Such an investigation

could help identify which visual attributes have a stronger influence to certain aesthetic impressions and which are more stable during time.

References

1. Lindgaard, G., Fernandes, G.J., Dudek, C., Brownet, J.: Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour and Information Technology* 25(2), 115–126 (2006)
2. Tractinsky, N., Cokhavi, A., Kirschenbaum, M., Sharfi, T.: Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human - Computer Studies* 64(11), 1071–1083 (2006)
3. Norman, D.A.: *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books, New York (2004)
4. Hassenzahl, M.: The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction* 19(4), 319–349 (2004)
5. Lavie, T., Tractinsky, N.: Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies* 60(3), 269–298 (2004)
6. Norman, D.A.: Introduction to this special section on beauty, goodness, and usability. *Human-Computer Interaction* 19(4), 311–318 (2004)
7. Fogg, B.G., Soohoo, C., Danielson, D., Marable, L., Stanford, J., Tauber, E.: How do people evaluate a Web site's credibility? Results from a large study. *Persuasive Technology Lab, Stanford University*, http://www.consumerwebwatch.org/news/report3_credibilityre-search/stanfordPTL_TOC.htm
8. Coates, D.: *Watches tell more than time: product design, information and the quest for elegance*. McGraw-Hill, London (2003)
9. Papachristos, E., Avouris, N.: The Subjective and Objective Nature of Website Aesthetic Impressions. In: Gross, T., et al. (eds.) *INTERACT 2009, Part I. LNCS*, vol. 5726, pp. 119–122. Springer, Heidelberg (2009)
10. Bassili, J.N.: *The how and why of response latency measurement in telephone surveys*. Jossey-Bass Publishers, San Francisco (1996)
11. Pham, M.T., Cohen, J.B., Pracejus, J.W., Hughes, G.D.: Affect monitoring and the primacy of feelings in judgment. *Journal of Consumer Research* 28, 167–188 (2001)

You Can Wear It, But Do They Want to Share It or Stare at It?

Arto Puikkonen¹, Anu Lehtio², and Antti Virolainen³

¹ Nokia Research Center, Visiokatu 1, 33720, Tampere, Finland

² Helsinki Institute for Information Technology HIIT, PO Box 68, 00014,
University of Helsinki, Finland

³ Nokia Research Center, Itämerenkatu 11-13, 00180, Helsinki, Finland
{arto.puikkonen, antti.virolainen}@nokia.com
{anu.lehtio}@helsinki.fi

Abstract. Wearable technologies are often used for supporting our daily lives instead of aiming to be entertaining. Yet it is in our daily lives that clothing is used to highlight our personas and engage others. In this paper, we describe what type of social acceptance issues might be worth to consider when it comes to entertaining and engaging wearable technology. Our user study with 10 participants was conducted by wearing a T-shirt that served as a display for an online game. The participants wore the T-shirt in their everyday surroundings. We gained a preliminary understanding on peoples' reactions and the suitability of this type of wearable technology for everyday usage. Our results indicate that established social boundaries for inappropriate attention influence the spectator experience with performative wearable technologies.

Keywords: Performative Wearable Devices, Social Interaction, Game Spectatorship.

1 Introduction

Rapid and continuous miniaturization of electronics and the fast evolution of new materials (e.g. nano materials) bring interesting possibilities for the use of wearable technologies. But despite that fact that wearable solutions prior have mainly dwelled in designs that aid us in our lives, the purchase of clothing is most often based on other factors than the level of aid received. Human actions are performative by nature and we are performing to the people surrounding us, to the spectators of everyday life [1]. Similarly, we choose the clothes we wear based on our own preferences, but also to either avoid or to catch the eye of our spectators. Wearable technologies are an interesting means to provide more elaborate and lively ways to support these needs.

We wanted to test a wearable concept that would support the wearers' desires of expressing self, but would also be entertaining for the spectators and maybe even entice social interaction. For the purpose, we designed a piece of clothing that was visually appealing, appeared to be interactive to enhance the spectator interest towards it and had a social component in the form a game. Our design was aimed to be

contrary to many other wearable technologies. We aimed heavily on the matter of visual appeal, instead of technological advances.

We chose to design a t-shirt that portrays an on-going game of Tic-Tac-Toe. The shirt was equipped with nine LED displays including a controller unit to control the displays. The purpose was to build the prototype in a way that the person wearing the shirt would not have to worry about the technology but wear it as any normal shirt. For the spectators the form of the T-shirt was supposed to look like a normal T-shirt.

To prevent the shirt becoming rigid and uncomfortable to wear, custom led displays were made. Discrete LEDs were assembled on a flexible substrate instead of using traditional rigid printed wiring boards. Wiring between the displays and a microcontroller was done with conductive strings. A normal T-shirt was used to hide the technology and give a nice finished look and feel while keeping the displays and wirings safe and sound between the two layers of textile (Fig. 1). Pre-programmed tic-tac-toe game software was stored on the memory of the micro controller and games were run continuously during the test period.

We formed our main research question based on the notion of spectator behavior: Does bringing this type of wearable technology into the context of peoples' everyday environment somehow overstep the social boundaries established for inappropriate attention? To this end we dug deeper into the subjective cognizance of the participants: How did they feel about wearing this type of wearable technology in their everyday environment and their descriptions and interpretations of the reactions and responses of people they came across while wearing the shirt.

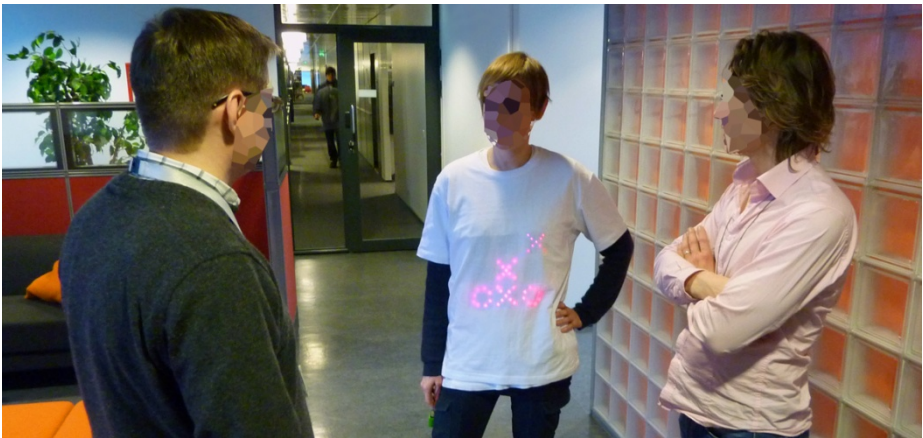


Fig. 1. Tic-Tac-Toe T-shirt in use displaying an on-going game

2 Related Work

The key concept contributing to our study from social sciences was ‘civil inattention’ introduced by Erving Goffman [2]. According to Goffman, social interaction is based on performative actions that take on dramaturgical forms. Bearing that in mind, civil inattention refers to a certain type of behavior that aims at maintaining some privacy

for the performer in surroundings that are otherwise public. This behavior consists of e.g. showing disinterest, appearing to be absent-minded and/ or acting oblivious as to what's happening. This type of behavior can typically be seen e.g. in urban setting where a lot of people inhabit a limited space. You politely "pretend" not to or make sure that you don't pay too much attention to what others are doing, saying etc.

In the context of games civil inattention is often not beneficial. Quite the opposite, immediate interaction around gaming forms a substantial part of the total experience. When features for social interaction around games are enabled, playing games is more fun and engaging [3, 4]. Additionally, Game cafes have been reported to garner their success from supporting conversation around games [5]. Friendships are built and maintained around gaming.

But interaction with the ones around you is not just about socializing. Reeves et al [6] have presented a taxonomy and resulting design principles for designing for the spectator. According to Reeves, interfaces should be approached from thinking them as secretive, expressive, magical or suspenseful. Each approach comes with its own benefits and concerns, but aims to create the best spectator experience. Around gaming, it has been reported that spectators most appreciate games designed to support visible actions and tactics, units in competitive play and emotions evoked during competition [7].

In addition to the multitude of technological problems related to wearable computing, social acceptability forms a noteworthy barrier. The bottom line being, as described by Rantanen et al [8], in addition to the technological features provided by a piece of clothing, it also has to provide the esthetical and functional properties one expects from clothing in general. Several issues related to wearable computing can be in opposition to these required properties, such as bulk [9] and dangling objects [10]. With wearable controllers even the placing of the controls has been reported to have an impact on the social acceptability. Holleis et al [10] found in their study that controllers on trousers, wristbands or separate bags received acceptance. On the contrary, the upper body, like a shirt or a scarf, was generally seen as not acceptable. Interestingly, Gabaret et al [11] found out that control has an effect on the social acceptability. The less control the wearer appeared to have over a wearable device, the more acceptable the device seemed to be to the ones around you. This notion was based on designing and comparing three different performative and artistic wearable computing devices.

The most interesting performative wearable devices in regards to our study are T-Qualizer [12] and BubbleBadge [13]. The T-Qualizer displays a music-related working equalizer on a t-shirt. The BubbleBadge is a wearable display designed to look like a brooch. As a brooch, it is meant to be placed on your chest, thus supporting face-to-face interaction.

3 User Study

To answer our research question, we planned a field trial with 10 participants. The age of the participants varied between 19-33, with 6 male and 4 female participants. Since we were interested in social acceptability issues that might arise if entertaining, wearable technology was introduced to workaday environment, we thought it best to

actually bring the technology to users. Therefore the study was conducted at a university campus with both students (6) and staff (4). All participants were recruited on the spot. Each of the participants was asked to wear our Tic-Tac-Toe T-shirt for a period of approximately one hour. Participants were encouraged to go about their daily duties as usual, not to let this experiment be in the way of their normal routines and not to avoid any of the customary situations that might occur. The length of one hour was chosen based on the rhythm of life at the campus. Participants were recruited during their breaks and the test period ended during the expected next break. Thus, the hour long cycle consisted of going somewhere, staying somewhere and coming back from somewhere. These places, called somewhere in this case, were classrooms, auditoriums, office spaces, cafes and lunch restaurants. A normal day at the campus constitutes of many of these cycles, and extending the experiment would have only resulted in multiple similar cycles. Also, due to the design of our t-shirt, more prolonged experiments were expected to make the t-shirt mundane in a relatively short period.

To get an idea on what was going on while the participants wore the t-shirt, we applied a light version of ethnographic methodology based on a form of participant observation [14]. To his aim, 5/10 participants were shadowed. With shadowing the focus was on getting insight on the reactions of both the “performer” and the “spectator”. In this case, the performer is the person wearing the t-shirt and the spectator anyone in the vicinity of the t-shirt. To get this insight, a moderator followed the user for the whole time the t-shirt was worn and took notes. Both the performer’s impressions and comments and the observed reactions of spectators were documented. 5/10 users were shadowed. The rest were expected to spend most of their time in areas that we couldn’t access due to drawing too much attention to ourselves. Dividing the users into two groups enabled us to evaluate, at least on a superficial level, if the shadowing had had any impact on the performers’ answers to the questionnaire. Naturally, during the shadowing the moderator tried to be as inconspicuous as possible.

During shadowing, the comments of the users were written down as well as the reactions and the comments from spectators. Since the users went about their business as they usually would and the moderator aimed not to be intrusive, we weren’t able to get all the comments made.

Participants weren’t explicitly encouraged or told to draw attention to the shirt. In the short brief we gave about the study, we informed the participants that the t-shirt was “a display” for an actual ongoing tic-tac-toe game played on the Internet. The participant would hence basically act as a display for it. The participants were given a mobile phone to give a more authentic feel of an actual online game. We told the participants that the mobile phone was downloading the game from the Internet. We chose to fabricate the story of an actual online game to support the feeling of actually being interactive and as the T-shirt’s game data in reality could benefit of being drawn from the Internet.

After returning the t-shirt to us, or when the shadowing moderator ended the session, the participants were asked to fill out a questionnaire. In a situation, where someone acted as a spectator for a longer period of time, the spectator was also asked to fill out a form that was especially drawn up for this type of a situation. The questionnaires consisted of both rating scales and open-ended questions about performer experience, spectator awareness, acceptability, interest towards the technology and areas of improvement. In addition, the participants gave rich verbal commentary after

the test session. These comments gave us knowledge that weren't present in the filled out questionnaire. In total, the combination of shadowing, questionnaires and verbal commenting gave us a thorough understanding of the results.

4 Results

The participants wore the t-shirt in varied locations, thus meeting a varied amount of possible spectators and varied level of direct interaction with others. Based on our observations the t-shirt was worn in the corridors, small classrooms, a large auditorium, small office rooms, open office spaces with cubicles and in cafes. The most common reactions the users met with while wearing the shirt were looks, (hidden) stares, smiles, laughter, and "raised eyebrows". The users didn't find this attention in the least disturbing. These overall reactions evoked by the shirt were quite subtle in nature and according to the users even surprisingly so. A comment made by one of the participants serves as an example on this: "*I was surprised, because it felt like people were paying less attention to me than usual.*" (User 7, Female) These notions by participants seem to suggest the presence of Goffman's [2] 'civil inattention', possibly even in an amplified form. It's socially acceptable to e.g. glance at a person you're passing by in a corridor. However, based on the observations, the glances by passers-by were very discrete and sometimes even missing altogether. Consequently, the observations made during shadowing support these participants' notions of somehow "being overlook".

The avoiding of eye contact, looking through someone etc. are typical forms of civil inattention. And this was the case in many situations. Our observations noted glances from a distance, but a lack of glances from shorter distances. It seemed that the spectators noticed the performer and that "something was up" from afar. So when they got closer, they were careful not to stare etc. In the context of civil inattention this could be interpreted as a sign of trying to assure that no embarrassment to one self or the performer would follow.

All the participants (6/10) that showed the shirt to someone did so only to people they already knew. The most common reasons for showing the shirt to someone were: the users thought the shirt was fun and/or wanted to see the reactions of their friends/co-workers/fellow students. Also, by drawing attention to the shirt themselves, the users reported feeling more in control of the situation. The users, who didn't show the shirt to anyone explicitly, explained this being due to not seeing anyone they knew. Since the users were instructed to go about their business as usual, a few of them reported it being useless or meaningless e.g. "*to sit in one's cubicle where there are no people around.*" (User 1, Female). This creates one limitation to the people's willingness to wear the shirt to places, where interaction around the shirt might not naturally occur. If there's no one to show the shirt to or no one to take a look at it, there seems to be little point in wearing it at all. This issue didn't come up with participants, who wore the shirt for their lunch or coffee breaks.

When compared to the findings with BubbleBadge [13], where the most fun was had by the spectators, it is interesting to realize the impact of civil inattention. It was clear from the users' responses that when attention was not received, the biggest factor for fun came from sharing. This sharing most arose in the form of showing the

shirt to friends, who are natural recipients for sharing. Additionally, as no active or forthright attention or actions exhibited by strangers were reported, it would seem that the social interaction between the spectator and performer was greatly influenced by social closeness instead of just the actual setting. During shadowing it was also observed that social aspects of hierarchy might play a part in the matter. *“Why did I turn on my heels? That’s a good question, I didn’t even think about it. Those are my bosses and if they saw with this shirt on I’m sure they’d start making jokes or something.”* (User 7, Female)

When the shirt was shown to somebody, the spectators were mainly interested in finding out how the shirt works. As the shirt was designed to be somewhat magical, based on the taxonomy of Reeves et al. [6], this question was to be expected. When the social situation allowed the question to be asked, the spectators naturally wanted to understand “the magic”. The spectators also asked what does it do and why the participant was wearing it. Two out of ten participants reported that someone had expressed interest in doing something with the shirt themselves - they either wanted to be able to play the game or wear the shirt. 8/10 participants wanted to be able to have some control over what’s happening on and with the shirt. Most common concern was that the possible content of the shirt might be something the participant would find meaningless or being opposed to personal taste or values. *“I’d like to decide what I’m displaying, yes. I wouldn’t want to e.g. advertise McDonalds or a band I really don’t like.”* (User 9, Male) Also the ability to turn the shirt on and off was mentioned. The tic-tac-toe game currently displayed on the shirt was seen as ok, but almost all of the users suspected that it wouldn’t engage spectators for long. However, this wasn’t seen as a big problem by the participants since the users didn’t expect anyone to look at the shirt *“for hours”* anyway (User 10, female).

When choosing to utilize Tic-Tac-Toe in the design, we acknowledged that one design would not perfectly suit all participants. Naturally, when choosing a piece of clothing to wear, the clothing is often chosen based on the most appealing design. Our decision to use Tic-Tac-Toe was influenced by wanting an interactive design that still could be about self-expression and be appealing. Based on the feedback from the participants, the idea behind the idea was appealing, but not the current design of it. The participants stated that in an everyday situation the design should be more refined. When evaluating the current appeal, the shirt received an average of 2,4/5. On the other hand, the participants reported rather a high willingness (avg. 4,1/5) to wear the shirt again, based on how it suited their social surroundings. Some users even expressed enjoying the possibility *“to be a little on view”* (User 3, female) as a motivation for wearing the shirt. These results combined confirm similar findings made by Rantanen [8]. That is, in addition to the appealing technological features, wearable technology has to provide the esthetical and functional properties one expects of clothing in general.

Even though the average for the willingness to wear the shirt suggests that the everyday surroundings didn’t present a problem for wearing the shirt, the answers to open-ended questions offer a corrective to this. The users listed suitable situations and places for using the shirt as follows: bar, party, for opening up a conversation, a night out, concert and nightclub. Work, business meeting and formal occasions were on the top of the list of the most inappropriate places to wear the shirt. On the other hand, for some participants the shirt *“gave on excuse”* to be on view. A few spectators also

brought up this matter, combining it with what they termed to be the personality of the user. *“It’s ok I guess, but for X (User 6) to be wearing that... it’s not really his style, his thing, if you know what I mean? It’s not his personality.”*

When talking about the idea of the t-shirt, the participants also stated understanding that the idea of being a display on the move requires that the information presented has to be short. The users suggested that in addition to games (Worms, Tetris, Super Mario, Chess) the shirt could display e.g. weather forecasts, advertisements, news and videos. Using the internet for content retrieving was seen as a good idea and facilitating many varied options for the content displayed.

5 Conclusions

In this paper we have presented a T-shirt that portrays an on-going game of Tic-Tac-Toe. With the help of the T-shirt, our goal was to understand the impact of civil inattention with wearable performative clothing. For this goal, we conducted a user study of 10 participants, who each wore the T-shirt for an hour in their everyday surroundings. To fully understand the different aspects of the matter, we dug deeper into the subjective cognizance of the participants with shadowing, questionnaires and discussions.

Our results indicate that civil inattention exists when using performative clothing. Our participants reported noticing some glances, but at the same time, noticing also an interesting lack of glances. Similarly our observations based on shadowing, support this notion. To compensate the lack of attention to something that the participants felt excited about, the participants chose to share this experience with their friends.

The participants also suggested acceptability towards utilizing wearable technology as a performative and eye-catching element. Our results were also in alignment with previous studies showing that wearable technologies need to provide aesthetic and functional properties, just like clothing in general. Based on our relatively small sample data, we cannot make deeper claims of the reception of such technologies, but we feel that this is a good starting point for further investigation related to social acceptability. And even the fact of agreeing to wear the shirt in the populated surroundings, where the users inhabit and work, can be considered to be at least a weak sign of acceptance for introducing wearable, personal technology for such surroundings. Since the study was the first done with our T-shirt, it acts primarily as an explorative study for testing and mapping out the social acceptability issues. The results highlight the acceptability issues that arise when introducing a new technology into the mix. To this end, the study setup provided excellent results. On the other hand, the results are mainly relevant to the introduction, not long-term use. To understand the long-term issues, more research needs to be done.

For future work, we aim to support and study features for richer interaction and enhanced spectator experience.

References

- [1] Rico, J., Jacucci, G., Reeves, S., Koefoed Hansen, L., Brewster, S.: Designing for Performative Interactions in Public Spaces. In: Proc. Ubicomp 2010, Copenhagen, Denmark (September 23-29, 2010)

- [2] Goffman, E.: *The Presentation of Self in Everyday Life*. Doubleday, Garden City (1959)
- [3] Gajadhar, B.J., de Kort, Y.A.W., Ijsselsteijn, W.A.: Shared funn is doubled fun: Player enjoyment as a function of social setting. In: Markopolous, P., de Ruyter, B., Ijsselsteijn, W., Rowland, D. (eds.) *Fun and Games 2008*. LNCS, vol. 5294, pp. 106–117. Springer, Heidelberg (2008)
- [4] Gajadhar, B.J., de Kort, Y.A.W., Ijsselsteijn, W.A.: Rules of Engagement: Influence of Co-Player Presence on Player Involvement in Digital Games. *International Journal of Gaming and Computer-Mediated Simulations* (2008) (in press)
- [5] Jonsson, F.: A Public Space of their own. A Fieldstudy of a Game Café as a Third Place. In: *Nordic DiGR*, Stockholm, Sweden (August 16-17, 2010)
- [6] Reeves, S., Benford, S., O'Malley, C., Fraser, M.: Designing the spectator experience. In: *Proceedings of CHI 2005*. ACM, New York (2005)
- [7] Cheung, G., Huang, J.: Starcraft from the Stands. In: *CHI 2011*, Vancouver, Canada (May 7-12, 2011)
- [8] Rantanen, J., Karinsalo, T., Mäkinen, M., Talvenmaa, P., Tasanen, M., Vanhala, J., Alfthan, N., Impiö, J., Malmivaara, M., Matala, R., Reho, A.: Smart Clothing for the Arctic Environment. In: *Proc. ISWC 2000*, p.15 (2000)
- [9] Dunne, L.E., Ashdown, S.P., McDonald, E.: 'Smart Systems': Wearable Integration of Intelligent Technology. In: *International Center for Excellence in Wearable Computing and Smart Fashion Products*, Cottbus, Germany (December 9-11, 2002)
- [10] Holleis, P., Schmidt, A., Paasovaara, S., Puikkonen, A., Häkkinen, J.: Evaluating capacitive touch input on clothes. In: ter Hofte, G.H., Mulder, I., de Ruyter, B.E.R. (eds.) *ACM International Conference Proceeding Series*, pp. 81–90. ACM, New York (2008)
- [11] Garabet, A., Mann, S., Fung, J.: Exploring Design through Wearable Computing Art(ifacts). In: *Extended Abstracts CHI 2002*, Minneapolis, Minnesota, pp. 634–635. ACM Press, New York (2002)
- [12] T-Qualizer, <http://www.tqualizer.com/>
- [13] Falk, J., Bjork, S.: The BubbleBadge: A Wearable Public Display. In: *Proceedings of CHI 1999*, pp. 318–319 (1999)
- [14] Chicago school (sociology) (April 7, 2011), [http://en.wikipedia.org/wiki/Chicago_school_\(sociology\)](http://en.wikipedia.org/wiki/Chicago_school_(sociology))

Design and Evaluation of Interaction Technology for Medical Team Meetings

Alex Olwal, Oscar Frykholm, Kristina Groth, and Jonas Moll

School of Computer Science and Communication
KTH (Royal Institute of Technology), Stockholm, Sweden
{alx, frykholm, kicki, jomol}@csc.kth.se

Abstract. Multi-disciplinary team meetings (MDTMs) are essential in health-care, where medical specialists discuss diagnosis and treatment of patients. We introduce a prototype multi-display groupware system, intended to augment the discussions of medical imagery, through a range of input mechanisms, multi-user interfaces and interaction techniques on multi-touch devices and pen-based technologies. Observations of MDTMs, as well as interviews and observations of surgeons and radiologists, serve as a foundation for guidelines and a set of implemented techniques. We present a detailed analysis of a study where the techniques' potential was explored with radiologists and surgeons of different specialties and varying expertise. The results show that the implemented technologies have the potential to bring numerous benefits to the team meetings with minimal modification to the current workflow. We discuss how they can augment the expressiveness and communication between meeting participants, facilitate understanding for novices, and improve remote collaboration.

Keywords: Medical team meetings, collaboration, single-display groupware, multi-display groupware, multi-touch, pen, mobile.

1 Introduction

Multi-disciplinary teams in modern healthcare are facilitating the discussion of patients, decisions on diagnosis and treatment, operation planning, interventions and surgery [5, 7, 9, 15, 10, 20, 6]. The medical specialists participating in the meetings are not always located in the same facilities, or even city, whereby tele- or video-conferencing is frequently used. During the meetings, specialists present relevant information about patient cases, based on their domain-specific expertise. Current meeting facilities typically only allow image presentation and demonstration by the radiologist and/or pathologist, whereas the other meeting participants have no means for interacting with the material [7, 9, 11, 5]. To support the changing style of collaboration, new, effective user interfaces are needed.

We have worked closely with a surgical department at a university hospital, which focuses on complicated deceases in the upper abdomen, where multi-disciplinary team meetings (MDTMs) are a central part of patient care. In this context, we are interested in applying and evaluating technology that can improve understanding,

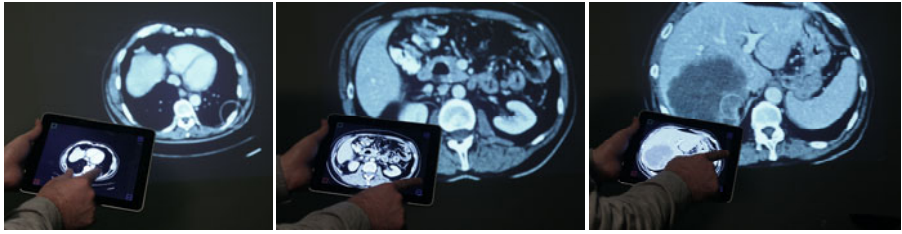


Fig. 1. Synchronized interaction with projected radiology material, using multi-touch gestures and widgets to zoom and scroll through the image stack

simplify group discussion and increase collaboration for safer and faster medical decisions [5]. We have therefore developed a set of tools and techniques, focusing on multi-display groupware (MDG), as shown in Figure 1. In the spirit of previous work [1, 13, 16, 19, 21], our goal has been to distribute the interactive capabilities at such meetings among participants. While our work is applied to the medical domain, we also find it generally applicable to other types of MDG and single-display groupware (SDG) scenarios [21, 13, 23]. To ensure qualitative feedback without influence from peers and to minimize the effect of time constraints, this study explores the collaborative techniques with one participant at a time in a simulated team meeting. This first evaluation of tools for facilitated collaboration in MDTMs will form the basis for a large-scale user testing in the real setting.

We first review related work, followed by an overview of the medical team meetings that we are studying and a discussion of a number of initial studies that we conducted, which form the basis for this work. We then present our method, a set of design guidelines and the implemented techniques. The main part focuses on our user study and results, followed by a Discussion, Conclusions and Future Work.

2 Background

Several researchers report on detailed field studies of MDTMs [e.g., 5, 6, 7, 8, 9, 11]. Time is an important aspect when judging the efficiency of the meeting [7, 5] and supporting technology must take this into account, as well as specific participant roles [6]. Although some problems regarding efficiency and effects on the group work process were likely to arise, Kane and Luz [7] acknowledge the need for technology for collaborative decision-making. Unsupported needs include the ability to point to areas of interest and to make annotations on shared displays [8]. Laser pointers were shown to partially alleviate these deficiencies, although it was noted that the laser dot could not be seen at remote locations during teleconferencing meetings [7].

Several projects have explored digital tools for multi-user interaction on single or multiple displays. The Pebbles Project [13, 14] is a collaboration framework for multiple handheld computers that are connected to a shared display, providing remote keyboard and mouse control, and multi-user drawing with individual cursors. M-Pad [19] uses tracked handhelds for advanced multi-user interaction techniques with whiteboards through toolglasses, palettes, cross-device transfer, and annotations. While these systems were pioneering, they had no wireless connectivity and their graphics performance limited possibilities for manipulating medical images.



Fig. 2. Left) Radiologist controls the interaction with the imagery from his workstation by navigating (scrolling through image stacks) and pointing out findings with the mouse. Right) Medical specialists from different disciplines (e.g., pathology, oncology and hepatology) view and discuss the imagery from a distance, and can only refer to areas of interest through verbal means or finger pointing.

Architectures for mobile interaction with shared displays [16] and multi-cursor interaction [1] illustrate the complexity when designing multi-display groupware environments. Tse et al. [22] describe a set of experiments where shared workspaces were partitioned by participants into individual areas to address interference, while one of the motivations for designing cooperative gestures [12] was to clarify users' intentions. Wallace et al. [23] discuss the results from a study comparing SDGs (multiple input devices, single display) with MDGs (multiple input devices, individual displays). They conclude that SDGs benefit in awareness of collaborators' activities, but may be distracting and can interfere with the group's primary tasks. MDGs, on the other hand, may provide a private, customizable, workspace with fewer distractions, but can instead require more efforts for coordinating team work.

Lee et al. [10] and Eng et al. [2] describe remote consultations where radiologists and physicians use a synchronized image viewer. The systems enable shared reviewing with joint image navigation, synchronized cursors for pointing, and image annotations. Shareable displays [18] have also been developed using PDAs to access and navigate radiology material on larger displays. Recent advances in mobile technology, allow us to overcome the limitations of previous systems [3] to support distributed interaction with radiology imagery.

Our paper builds on a body of work, in which we have studied a large number of MDTMs at a gastro-surgical department over four years. Results from these studies indicate many benefits if also participants in these meetings could navigate the material, point to regions of interest, and annotate important features.

2.1 MDTMs at Gastro

The centralization of specialized medical care to regional university hospitals is a recent strategy for quality improvement and effectiveness. The gastro-surgical department (Gastro) that we have studied has the regional responsibility for specialised care in the upper part of the abdomen (i.e., complex diseases in the liver, pancreas and esophagus). The 25 surgeons collaborate closely with several other units within the university hospital organisation, including pathology, radiology, oncology,

and hepatology. The weekly MDTMs (See Figure 2) are an important element of the work process at Gastro.

The patient care pathway at Gastro [4] consists of four main steps: coordination of medical material, decision on diagnosis and treatment, surgical treatment, and post-operative treatment. Gastro regularly uses three types of MDTMs during the patient care pathway.

The *Decision Meeting* is a forum for discussion of patient diagnosis, operability and resectability, to decide on the next steps in the care, such as surgery, radiology, oncology treatment, or the need for additional examinations. There are a number of key roles at the meeting. First, a surgeon summarizes the patient's medical history and formalises the discussion topic. Second, a radiologist, seated at a workstation with three high-resolution displays, presents the radiological analysis, while participants follow the walk-through of the medical imagery on two projection screens. Third, if there are pathology reports, then these are presented by a pathologist. Fourth, a senior surgeon leads the following discussion, which relies on the display and navigation of medical imagery, to reach a consensus decision. A group of medical specialists from different disciplines participate with comments, questions and discussions. Less experienced medical doctors attend as part of their training, but do not actively participate in the discussions. The meetings must be streamlined and efficient due to a high number of patients to discuss by the many attendees under time constraints.

Pre-operative Meetings (Pre-ops) are similar to Decision Meetings, but have fewer participants, are more focused on operation strategy and planning and have no time constraints. They are typically attended by three to six surgeons, one radiologist and occasionally one pathologist, all from the same hospital, and take place the week before surgery. The video link is not used, but specialists from other departments can participate if needed.

Post-operative Meetings (Post-ops) allow pre-op attendees to share their experiences after surgery to, for example, discuss how well the pre-op planning matched the surgical procedures. These are only conducted for pancreas cases.

2.2 Initial Studies at Gastro

Our work is based on four years of field studies where data has been collected in an on-going process of ethnographic fieldwork and participatory design projects. Our early MDTM observations have showed a need, but limited possibilities, for interaction with patient information. To better understand the interaction with radiology imagery during MDTMs, we observed and analysed twelve additional Decision Meetings and Pre-ops during 2009. The meetings were video recorded and the sections where surgeons specifically tried to point to and navigate in the images were analysed in detail. Similar to related work [15], it was observed that the radiologist relies heavily on mouse pointing when presenting and explaining the medical imagery. The analysis of the observations show that also the surgeons point to parts in the medical imagery (from their seats or by approaching the screen) to illustrate how operations should be performed or when asking for clarifications. This indicates a need to illustrate a procedure or make a deictic reference to a detail in the imagery. It was also clear that the radiologist, surgeons and other specialists have a formalized language with an established common ground that make it easier to

verbally refer to parts in the imagery during their discussions. We also observed occasions where some participants used laser pointers to clarify their intentions.

To further understand the interaction during MDTMs, a follow-up field study of three consecutive Pre-ops (each lasting ~40 min) was conducted in which laser pointers were handed out to all participants, except the radiologist. The objective was to observe how laser pointer usage would affect meeting discussions. Participants were not given specific instructions on how or when to use the laser pointers. In the analysis, we found that the laser pointers were used:

- 1) To point and “draw” in different parts of the projected imagery
- 2) For joint navigation and guidance
- 3) To help coordinate communication by pointing to smaller features while asking for more detailed information
- 4) For gestures that allowed surgeons to illustrate plans, e.g., for cuts of organs

The setup’s major drawback was that the radiologist had to direct his attention from the workstation to see the laser dot on the projection screens, preventing simultaneous control of the workstation and maintained awareness of the surgeons’ pointing.

The advantage of laser pointers, from a technological standpoint, is low cost, portability and infrastructural independence. Still, despite the benefits and immediate availability, laser pointers have not been widely adopted by participants at the MDTMs. From observations it became clear that laser pointers do not meet the requirements of multi-display meetings, where radiologists, surgeons, and remote participants are viewing three separate, but digitally synchronized displays. A surgeon’s laser pointer on the projection will thus not be visible on the radiologist’s workstation or through the videoconferencing system, which turns out to be prohibitive for practical use. Also, the laser pointer study showed that the person pointing is using words like “here” instead of verbally explaining a part in the image, which is making the referencing even more ambiguous for participants viewing other displays. While it would be possible to track and distribute laser dots as digital cursors using cameras [17], the required infrastructural modifications inspired us to instead develop techniques that would support more advanced and flexible multi-user interaction through digital tools, without the need to install any hardware in the MDTM room.

3 Prototypes for Collaborative Interaction Techniques

The objective of the study was to explore the functionality that new interaction techniques and devices can enable for MDTMs, and evaluate their potential for improving understanding, group discussion, collaboration and decision making. Data from our initial studies allowed us to make informed design decisions for the candidate interaction technology to be used in the study. We first decided to prioritize the most important subset of medical imaging functionality from the radiologist’s workstation, based on the observed needs of the meeting participants in our studies:

- 1) Shared/collaborative pointing
- 2) Navigation (zooming/panning/scrolling in image stack)
- 3) Annotation (sketching/drawing)

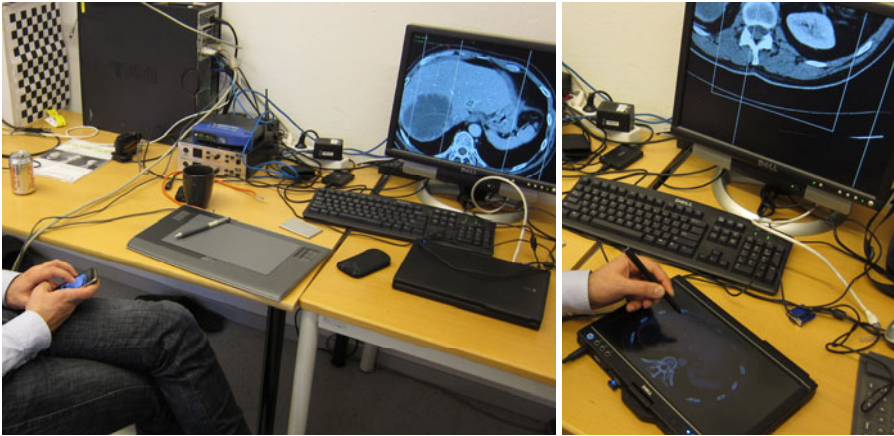


Fig. 3. Surgeons and radiologists participated in interactive prototyping sessions. Left) A radiologist manipulates content on a shared display using a mobile device. Right) Interaction on a Tablet PC (Dell XT2) that supports both pen and multi-touch input.

We interviewed and observed two senior radiologists at their hospital workstations (each ~2 h) to gain insight into workflow, important functionality, frequent operations and desired features. We then conducted an interactive workshop with a senior pancreas surgeon (~3 h) where we discussed the use of digital devices for interacting with the radiology images and related these ideas to typical MDTM discussions. A first version of the prototypes was developed after these sessions. The prototypes were demonstrated and discussed with the surgeon at a second workshop (also ~3 h), which provided additional feedback. Two interactive prototyping sessions (each ~2 h) were then conducted individually with one of the previously observed radiologists and one new radiologist. Similarly to the second workshop with the surgeon, they individually explored the first version of the prototype interfaces, but also answered questions regarding current practices/problems, potential for the proposed techniques and suggested functionality and features (See Figure 3). Audio and video was recorded for the sessions with the surgeons, and informal notes were taken during the radiologist sessions.

Insights from these sessions supported our preliminary ideas, inspired topics for the user study questionnaire and indicated an interesting potential for collaborative interaction techniques in MDTMs.

3.1 Design Guidelines

A set of design guidelines for the tools that we were developing for Pre-ops was identified from our initial studies and prototyping sessions:

Minimal changes to current meeting situation: The physicians are in general satisfied with routines and configuration of current meetings, and that the use of technology is limited to the radiology workstation. We thus wanted our tools to augment the meetings with minimal modifications to current workflow.

Sporadic interaction with minimal learning time: The users should be able to quickly access the needed functionality, as the dynamic meetings cannot afford complex and time-demanding user interfaces.

Arbitrary number of participants: The system should be able to support interaction from any user through an inherently scalable infrastructure.

Radiologist delegates control: As the radiologist is an expert of interpreting and navigating the medical imagery, and is in control of these processes today, it seems important to maintain that role. While many surgeons will likely want to interact with the imagery, this is not their area of expertise, and may not necessarily be effective. The interviews indicate that distributed interaction could be adopted as long as the radiologist could delegate control when needed.

Synchronization across displays: Current methods (e.g., finger pointing, walking up to the projection screens and the use of laser pointers) exclude remote participants and the radiologist. Digital tools, on the other hand, make it possible to synchronize interaction across multiple displays, which could especially be used to empower remote participants.

3.2 Interaction Techniques and Implementation

The design guidelines informed the choice of the technology and architecture for the system. Thus, our system, for example, supports a radiologist's delegation of control, remote participation, and many other features. It is based on a PC server that projects medical imagery, while connected mobile devices are synchronized using the network, such that multiple users can interact using their own device.

Software architecture: Our framework is implemented using C++ and openFrameworks, which allows us to support most of the implemented functionality across multiple platforms (Linux, Mac OS X, Microsoft Windows XP-7 and Apple iOS). The mobile devices are synchronized with the server using OSC (OpenSoundControl), a UDP-based protocol for efficient, platform-independent data streaming.

Pen-based devices: We implemented our techniques on Wacom tablets (no display) and Tablet PCs (Dell XT2) that use electromagnetic pens with high precision/pressure and physical buttons, as shown in Figure 4.

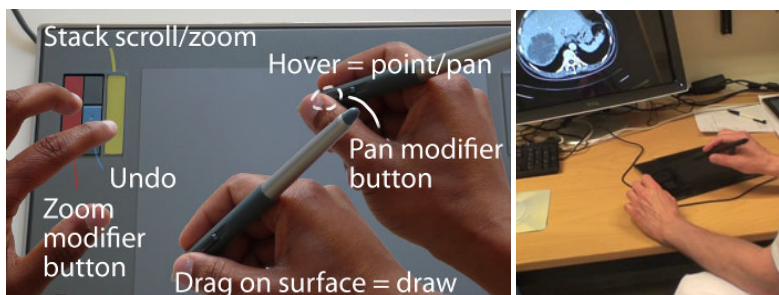


Fig. 4. The tablet uses buttons and a touch strip with a tracked pen for panning, zooming and drawing

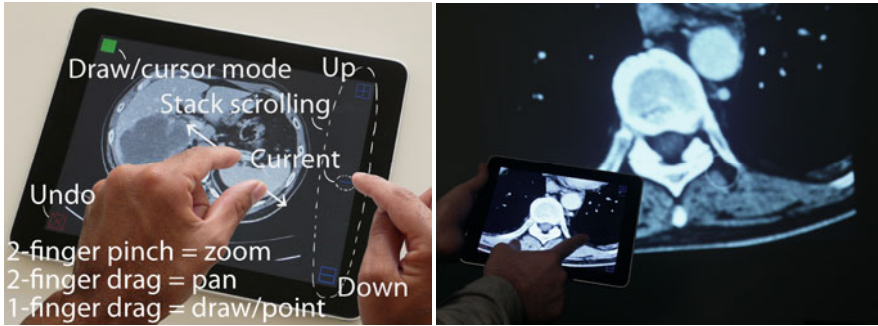


Fig. 5. The multi-touch displays use on-screen buttons and widgets in combination with multi-touch gestures

Multi-touch devices: We use Apple iOS devices (3.5" iPod Touch 3rd Generation and 9.7" iPad) as portable multi-touch displays (See Figure 5), and support larger screens through the multi-touch capabilities in Windows 7 (e.g., 21.5" Dell SX2210T, 22" 3M M2256PW, and 12" Dell XT2).

All device types were explored in the prototyping sessions with the surgeon and radiologists, and based on our observations and their feedback we chose to narrow it down to three devices for the user study: the screen-less tablet, and the small and large multi-touch devices (iPod Touch and iPad). We carefully selected the techniques through our initial prototyping and user observations, based on observed gestures at MDTMs and other studies. The hardware was chosen based on commercial availability, mobility, the possibility to use it outside MDTMs, and the different form factors. The pen-based device was selected as the lack of display avoids problem with private focus.

We developed three general interaction techniques for multi-user interaction.

Pointing: One of the primary motivations of the system is to enable participants to point accurately from any location in the room, or over the video conferencing link. It is achieved by touching and dragging with a single finger on the touch-screen devices, or by hovering the pen over the stylus-driven tablet. A coloured cursor distinguishes different users.

Navigation: A central component of the radiology material is composed of CT (computed tomography) or MRI (magnetic resonance imaging) image stacks, where the 3D volume of the human body and organs are viewed one 2D slice at a time. Here, zooming, panning and scrolling up/down in the stack are the most important means for navigation. Pan is controlled by dragging two fingers in the multi-touch interfaces, and zoom level is adjusted with two-finger pinching. A widget on the right-hand side visualizes the stack, with absolute access to each slice, and two buttons allow stepping up/down one slice at a time. On the pen-based device, the user pans by pressing a side button on the pen and hovering it over the tablet surface. Stack scrolling is controlled by sliding with a finger of the non-dominant hand on the tablet's touch-sensitive strip, while the simultaneous pressing of a device button zooms instead.

Annotation: Annotation with individual colours on each slice was implemented to support the exploration of temporary or permanent freehand annotations during the meetings. The multi-touch interface uses an on-screen button to toggle between pointing and drawing, whereas another button allows the user to undo the last stroke for the current slice. Dragging the pen on the tablet surface creates annotations that the user can undo using a dedicated button.

4 User Study Setup

We designed a formal qualitative user study to evaluate our interfaces and explore our target group's attitudes towards using digital tools for distributed pointing, navigation and annotation during MDTMs. The respondents explored pen-based interaction (Wacom tablet) and multi-touch interaction on a small and a large portable display (iPod Touch and iPad) in a simulated Pre-op setup where interaction was synchronized with a large display (see Figure 6). The user study was conducted with one physician at a time, with the study leader acting as radiologist, in control of the large display (which represented the projected screens used in the MDTMs). Two authentic patient cases (a liver and a pancreas case) were prepared in advance, including radiology images and text from the patient record (i.e., the radiology documentation from the Pre-op).

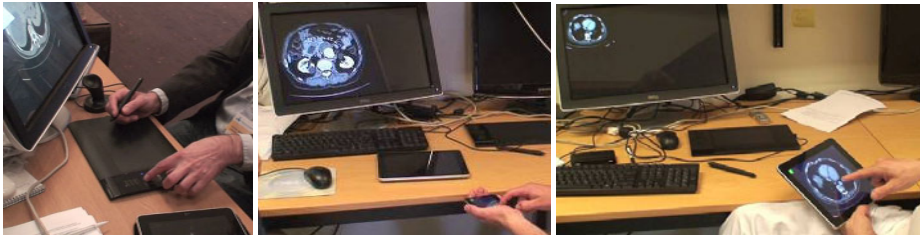


Fig. 6. Photographs from our user study, where participants explored the pen-based device (left), a small (middle) and a large (right) multi-touch display. The monitor represents the shared projection screen at the MDTMs.

We report on the results from individual sessions with eleven surgeons (one female) and one senior radiologist from the hospital. Seven of the surgeons were senior gastro surgeons specialised on a specific organ or intervention technique, four were consultants, four were gastro surgeons, one was a resident and one was a visiting gastro surgeon. The four consultants would typically lead the discussions, and the surgeons or the resident would present the case at the MDTMs. The radiologist was asked to respond to the study as if he was attending an MDTM where another radiologist demonstrated the material. The choice of focusing on surgeons is based on our field data showing that they are usually the ones asking more specific questions about the demonstrated imagery. We will broaden our studies to include more disciplines in the future. While the respondents lack experience in using ICT at these meetings, and access to technology is usually limited to PCs and medical devices, most of them were interested in exploring new technology that could support their

work. Most of them had used or were familiar with mobile touch-screen devices, knew about iPads but had not used one or seen one in real life, and were not familiar with Wacom Tablets. The oldest respondent was known, among the surgeons, to be conservative about new technology. The respondents were 35–62 years old (avg. 42 years) and received no compensation. The sessions lasted 18–52 min (avg. 36 min).

The study's objective was to identify and assess the potential of the implemented interfaces and technologies, and the type of functionality that would be useful at MDTMs. Questions from the study leader accompanied a pre-test questionnaire to collect initial attitudes. The respondents then explored the three interfaces (in randomized order) using think-aloud, combined with interview questions during the test, and a post-test questionnaire, where the usefulness of navigation, pointing and annotation in the different meeting situations, were respectively ranked on six-point Likert scales. Questionnaire data was not analysed statistically due to the number of respondents.

The respondents were asked to primarily consider the technology for Pre-op meetings, but the concluding discussion also concerned its applicability in other meetings or situations. The respondents were encouraged to comment and discuss with the study leader throughout the whole session. Sessions were recorded on video, transcribed and analysed in detail by categorising the data based on the current situation, potential risks, situations in which the devices could be used, and feedback on the interaction techniques and the input devices.

While individual sessions cannot fully simulate the multi-user interaction at the MDTMs, they were intended to allow us to gather focused feedback from each respondent without time constraints. The most senior participants at the MDTMs tend to dominate discussions given the formal experience-based hierarchies that dictate meeting roles. We thus balance our aim for a focused session and qualitative feedback from individuals with varying expertise and background, by having the interviewer take the role of the radiologist (controlling imagery from the workstation) with the participant simultaneously interacting using different devices. All respondents are well familiar with the MDTMs, and all senior surgeons and the radiologist attend every week. Two of the respondents also participated in the earlier study with laser pointers.

5 Results

The main target of the interaction in today's MDTMs is the radiology presentation and images. The radiologist is in control of the presentation and the other participants have to rely on other means of interacting with the images, for example when clarifications or more details are needed. In the pre-test discussion and questionnaire, almost all respondents emphasized the need for interaction with the images, especially pointing: *"Some colleagues have laser pointers ... They use them to point ... Several approach and point in the projection"* (resident F). However, some respondents stressed that Pre-ops are more suitable for interaction than Decision Meetings due to the large number of cases presented there.

5.1 Interaction Techniques: Pointing, Navigation and Annotation

Of the three interaction techniques, the respondents were most excited about pointing, which could help resolve ambiguous references, avoid misunderstanding, and improve communication for local and remote users: *“Pointing at a detail ... is very important”* (senior surgeon U). Four physicians said that they today rely on verbal explanation using common terminology (e.g., the eight liver segments can be referenced by index), point with their finger towards the projected imagery, or walk up to the projection screens to point. Some respondents found this acceptable and not a problem that they had reflected on as they considered verbal references easy to understand for the radiologist: *“You just say back a little ... scroll some”* (senior surgeon B) and *“I say that I want to see the blood vessel there”* (senior surgeon A). The ineffectiveness of verbal references was pointed out by several respondents who found it time-consuming to explain to the radiologist which area they would like them to navigate to. They mentioned that it was an abstract way of pointing and could cause misunderstandings due to the difficulty of verbally directing the radiologist, and that participants may thus hesitate to ask for clarifications.

Precise and unambiguous pointing would also make the discussion easier to follow for surgeons in training. Senior physicians tend to have developed a more precise language and can therefore refer more accurately to specific parts of the images, while it is typically more difficult for junior surgeons to interpret and follow the discussion: *“The experienced colleagues can use a completely different language, they are more precise than I can handle ... often they point, sort of sweeping ... but it’s not really obvious what they point to”* (resident F). The possibility for clearer pointing would, according to the resident, make it easier to follow and learn from the discussion. Thus, although experienced surgeons may be able to give accurate descriptions of areas of interest without pointing devices, their use of such devices might still be a great aid to less experienced surgeons.

Feedback on navigation primarily focused on image stack scrolling, while zooming and panning was emphasized for usability on the small display. Many respondents were not convinced about the need for navigation during MDTMs as they considered it the radiologist’s job: *“The radiologist presents the material very thoroughly and can see much better than I”* (surgeon C).

Annotations seemed less important for the Decision Meeting discussions, but numerous comments indicate potential benefit in other situations. As for precise pointing, referencing would be clearer for, e.g., residents and junior surgeons, when lesions and relations to arteries can be highlighted. Basic functionality, which radiologists have access to today, such as marking and measurement tools, were frequently asked for by several respondents. Also, one surgeon commented that drawn annotations could be useful in specific cases that need more detailed discussion.

Some surgeons emphasized the ability to save annotations: *“There is no point in marking temporarily — you’d want to save and continue later”* (surgeon T) and *“You’d mark the area that is unclear after the radiology presentation. The tumour often grows into a blood vessel ... you could actually mark it and refer to it here [while using the iPod]”* (surgeon S). This would help the surgeons understand patient cases faster if they could not attend the Pre-op, or wanted to refresh their memory in preparation for an MDTM or a surgical procedure. Simple symbols like arrows, labels

and numbers could complement the documentation, but also be useful for lectures and educational material. For a better overview of the annotations in radiology examinations (where there are multiple annotations in the image stack), an index of annotations, highlighting of annotated slices, and linked bookmarks to annotations were suggested.

5.2 Private Interaction vs. Shared Focus

Several respondents were excited to interact on their own high-quality display with easy-to-use controls. The large multi-touch device was appreciated for its large and clear display. Some respondents disliked the smaller device, while some thought that it was sufficient in size: *“I can see well on this one too, the zoom is so powerful”* (surgeon J). Some respondents expressed interest in exploring the imagery on their own and requesting control upon finding something they wanted to share: *“You could work maybe a little in parallel, not having to disturb everyone”* (senior surgeon B) and *“The radiologist is presenting, and you have your iPhone to see what he is presenting ... then you see something important, and you can annotate it on your own [device]”* (surgeon T). Still, it was generally considered that the radiologist should be in control of image manipulation on the shared display. If there was something respondents wanted to ask about or bring the group’s attention to, they could navigate and annotate on their personal device and ask the radiologist for permission to show it on the shared display.

Several respondents were, however, concerned that users would only focus on their private screen — a disadvantage compared to the screen-less tablet: *“The objective of the meeting is to have shared discussions ... if everyone is looking at their own [screen], then it isn’t a meeting”* (surgeon J), and *“It would be bad for the group discussion [if you had your own]”* (senior surgeon M). Also the radiologist commented on this: *“I may lose their attention when I demonstrate details ... they might not even hear what I say — they would be into their own analysis ... it is better that everybody has the same focus”*.

5.3 Applicability for Different Scenarios

Many respondents pointed out that the evaluated tools would be much more valuable at Pre-ops as the discussions in these meetings are more focused, the groups are smaller, there are fewer patients to discuss and the meetings are not under time pressure, in contrast to Decision Meetings. One surgeon said: *“There cannot be anything that disturbs the radiologist [at the Decision Meetings]. If you lose five minutes per patient, and have 19 patients ... That doesn’t work”* (surgeon T). Post-ops, on the other hand, are both less common (pancreas cases only) and less critical, as no decisions are taken. The respondents also commented on making annotations before the meeting, in order to save time during the meeting: *“[It] can be good to mark which structure is which — this is actually truncus hiliacus ... and that can be done in advance ... it is also how much time it [annotating the images] will take from the meeting”* (surgeon J).

The benefits of clarity and accessibility were, in particular, highlighted by the less experienced surgeons. The resident said that less experienced surgeons would be able

to follow the discussion better and would not need to disturb the more experienced surgeons if they can not follow what is referred to.

During the user studies, respondents also came to think of situations during the entire patient care pathway where the devices could be used. In team-based care, where multiple physicians share the responsibility for patients, it is important to be able to hand over patient information such that another physician can easily catch up on a case and quickly be able to make decisions. Decision Meeting protocols should accompany the imagery and examinations on to the Pre-op, where the surgical plan is annotated on the images. The plan could then be used when preparing the surgical procedure, especially if the surgeon in charge of the operation was not present at the Pre-op: *“The image information is very important ... today it [the surgical planning] is based on the Pre-op and that the surgeon is present to see and get a mental image for making a strategy to remember”* (surgeon J). The plan could be shown both to the patient and to the team in the operating room, and in the end ideally all the material could be accessible for presentation at a Post-op or follow-up meeting: *“It would be elegant if s/he [the surgeon] could display the images: ‘This is what we planned’, ‘This is what we did ... but when we physically got there ... this structure was there, and we decided to go here’ [simultaneously pointing in the image]”* (resident F). Some respondents found other situations more suitable than the MDTMs: *“I think this is fantastic [while using the iPod] ... then you can save, store and work with it on your own computer before the operation. ... The radiologists could annotate important things and I’d look at it on my computer later on. That would be more useful than if I annotate”* (surgeon T).

5.4 Interaction Technology: Multi-touch, Pens, Display, Size, Precision and Portability

Although respondents only used the three devices for about ten minutes each, most were able to give feedback on usability and ergonomics. On the other hand, the respondents were not always consistent in their preferences, as they might have preferred different devices depending on situation and purpose.

The same user interface and functionality was implemented for the two multi-touch devices. The smaller device was appreciated for its portable size, but at the same time several respondents found the display too small for efficient use. Numerous respondents, when interacting with it, only looked at the large, shared display. One respondent even stated that the small display was not needed, as it was sufficient to control the content on the shared display with multi-touch input. The zoom function of the multi-touch devices was much appreciated, as well as the responsive interface and display quality. The large device inherits most of the perceived positive properties from the smaller, but its size accounted for better precision and overview, and was therefore generally more appreciated.

Initially, the pen-based tablet was perceived as complicated to use since the functionality relied on two buttons, a scroll wheel and the pen: *“I can tell you directly that this seems awkward”* (senior surgeon U). On the other hand, when the respondents had used it for a few minutes, seven of them appreciated the device primarily for its high precision when pointing and drawing, and due to the fact that the large display must be viewed (which during a meeting would make participants

focus on the same display): “*Once you get used to it, it seems quite user-friendly, in my opinion ... this is good enough for drawing*” (surgeon J).

Two respondents did not prioritize drawing precision: “*Precise sketching is not that important because I think sketches and annotations are only important when referring to the operating surgeon that you haven’t missed anything, that you haven’t missed any point that was discussed*” (surgeon S) and “*The sketches would be quick, you wouldn’t have to be very detailed [while using the Wacom]*” (surgeon J). Three respondents were disappointed that they were not able to also use a stylus on the touch-screen devices. Two respondents believed that a high-precision device would probably be more relevant for the radiologist and the chair of the meeting, since other participants would probably not need the same level of precision. Thus, according to those participants, not everyone in the meeting needs to be able to interact with the medical imagery through networked devices.

Seven respondents were impressed by the interface and portability of the smaller multi-touch device while the larger was considered, by four respondents, too heavy (680 g) and large (1.34×19×24.3 cm) to carry around. Several respondents described ways in which they could be more effective if they had access to portable technology: benefits during individual preparation (e.g., at home, the evening before surgery) and for patient consultation. One of the surgeons said “*I would look briefly right before surgery*” (senior surgeon B), and another one said “*... in a hospital round ... [to patient] here is the problem, and we will solve it by removing this part*” (Surgeon S).

5.5 Suggested Improvements, Additional Functionality and Potential Risks

It should be stressed that most respondents believed that a limited set of basic and easy-to-use tools (for pointing, navigation and annotation) was sufficient. However, the respondents did also suggest additional functionality that could support other situations in their work. As all respondents, except one, were surgeons, they did not have access to a radiologist’s set of tools, and some of them mentioned features such as copying or overlaying annotations over an image sequence, comparing marked-out lesions over different image slices or examinations, and splitting the screen into two or four viewports.

Text annotations were frequently suggested as support for, and clarification of, discussed findings in the images. One surgeon said “*Put the number one here, and write what we plan to do*” (resident F). These text annotations could be combined with standardized options, accessible through, e.g., a drop-down menu, for simplicity and consistency. Some respondents even speculated about replacing the Pre-op meeting protocol with a set of text-annotated images as the predefined options could mandate a standard for the required information. Annotations from the meetings should of course be linked to the medical record, and physicians should at any time be able to access them.

Due to the way pointing was implemented on the devices (a circle with a size relative to the different display sizes) several respondents wanted a more precise pointing tool, such as an arrow.

While almost all respondents were enthusiastic about the possibilities of using interaction tools, a few concerns were raised about a multi-user system. Several surgeons commented that the radiologists are currently in control of demonstrating the

images and that each speciality should do what they are best at. Most respondents were concerned about “chaos” if multiple people tried to interact simultaneously with the radiology images. Several respondents suggested that this would probably be resolved through social protocols: “*You’d have to take turns, just like when we speak*” (surgeon J). Some respondents suggested that the radiologist should delegate control when needed. Another suggestion was that maybe not all participants should have devices, but only a few senior persons.

6 Discussion

Our initial studies, as well as field studies by other researchers (e.g. [7, 8]), show that surgeons often point to certain areas in medical images during MDTMs. The results from these earlier studies, and interview results from this study, suggest that using gestures far away from the projected images is neither sufficiently clear nor precise. Technical support for pointing and annotating seem to be able to alleviate such problems. Although not tested for statistical significance, our qualitative data gives strong indications for the potential of such interaction technology.

Of the three interaction techniques that we evaluated, pointing was considered most important. Our qualitative data highlighted the potential for precise pointing with digital tools that are synchronized across multiple displays, as it may save time in the discussion, include remote sites and make references more explicit for less experienced participants. Varying level of pointing precision should be supported on all devices, as a physician might want to refer to both small details and larger areas in an image. This would benefit discussions by minimizing ambiguous referencing. Annotation can, on the other hand, be accomplished with lower precision. It is not important to, for instance, exactly outline a lesion, or draw the exact path for incision during pre-operative planning. It is sufficient to highlight the area in which a lesion is located, or sketch the hypothetical incision, as the exact incision can only be determined during surgery.

Annotations could help clarify the discussion further, but were considered perhaps even more interesting when saved with the imagery for use in other situations. As anticipated, these techniques were confirmed to be most useful for Pre-ops, whereas the system should allow the data to be reused along the patient care pathway, such as for individual preparation before surgery or in Post-ops. Annotations made during the meetings, both in the form of sketches or, as several physicians suggested, text, should be saved for future use. As the physicians’ work is team-based and several physicians access the patient records during the patient care pathway, annotations would help information handovers. Radiology images augmented with additional information would make it easier and faster to understand what was discussed in a meeting, overview the most important images and findings, and review proposed surgical strategies.

While most study participants felt comfortable with the radiologist navigating the material, implicit navigation on the devices is obviously necessary in the user interface to support interaction with areas of interest.

Concerns were raised about issues that, for example, could arise with personal displays that distract from the discussion, or if multiple users attempted simultaneous

interaction. Thus, social protocols and technical policies that control the use need to be explored and evaluated before implementation.

Kane and Luz also discuss new technology's effect on group dynamics [8]. Often, participants already have a pre-defined role during the MDTMs, which would be affected by the introduction of tools that delegate control to other participants, an issue mentioned by several of our study participants. There was, e.g., concern that senior surgeons might start dominating MDTMs with less experienced radiologists. Surgeons and radiologists in our study, however, thought that communication protocols for turn taking would develop in the same way as when they are talking.

The three devices had their respective advantages and disadvantages, such as the pen's precision vs. the multi-touch interaction that was considered more intuitive, the ergonomic advantage of a large display size vs. portability, or the existence of a display vs. the potential risk for distraction from the shared screen. Whether, for example, personal navigation should be implemented during MDTMs, or if only a shared view should be used was also an issue identified by Wallace et al. [23], and our next step is to explore multi-user interaction during real MDTMs to evaluate how this technology affects group dynamics and to assess the importance of personal interaction. We emphasize that it is not the comparison of the devices themselves that is valuable in this context, but the qualities the study participants appreciated or found relevant.

Interestingly, several surgeons spontaneously suggested additional activities and scenarios in which the technology could be used, as there are many activities besides MDTMs in the patient care pathway. Surgery planning and preparation (e.g., the morning before the surgery) and patient consultation, were examples of other scenarios where participants thought the interactive devices could be useful. This is in accordance with previous research [4], which shows that cooperative and participatory design activities made them reflect on their own work and processes.

Physicians are also becoming increasingly mobile and will need access to updated, relevant patient information in different situations. Mobile devices could support a number of such situations, from hallway discussions to post-op reviews of complete cases. Personal navigation in a surgeon's office, for example, could be useful to understand anatomy and lesions for a specific case. Before performing surgery, the surgeon sometimes examines a 3D reconstruction of the volume. It is, in general, important for the surgeons to follow anatomical features (such as blood vessels, organs and lesions), to see how they intersect, for example. This is a functionality they all need, independent of activity (MDTMs, personal navigation or hallway discussions).

7 Conclusions and Future Work

We have developed and evaluated interaction technology for supporting emerging needs in multi-disciplinary collaboration for specialized medical care. Interviews, observations and a qualitative user study with surgeons and radiologists from different disciplines and varying levels of expertise, provided us with interesting insights for the next generation of digital tools for MDTMs. Unsurprisingly; the reactions were

overall positive, as our developed system clearly demonstrated the potential benefits of new interactive capabilities.

Our interaction techniques (pointing, navigation and annotation) for multi-user interaction were implemented on two multi-touch devices (small and large) and a pen-based tablet. The use of laser pointers that have been observed by several researchers [7, 15] does not support the typical multi-site video-mediated MDTMs. Digital tools do, in contrast, address the problems caused by, e.g., laser pointers that are not visible to all collaborating participants, without requiring additional hardware installation (e.g., cameras for laser pointer tracking [17]). While the proposed technology for augmenting the meetings is advanced, we emphasize the importance of the perceived simplicity of the interaction techniques. We believe that three classes of devices are relevant for further exploration at MDTMs: Simple pointing tools (e.g., multiple wireless mice or trackpads), portable devices (e.g., mobile devices with large touch screens) and larger wireless devices with high-resolution displays that support both stylus (precision) and touch input (e.g., Tablet PCs).

We build on previous work and field studies, by introducing technology and strategies for interacting with information presented during MDTMs and provide a set of design guidelines, the importance of which were implicitly confirmed in the study, that we hope will serve as inspiration for future MDTM systems. Our results show that more precise referencing could augment the expressiveness of participants, improve the communication with the other participants (including the radiologist), help less experienced participants follow the discussion, and bridge the gap to remote experts.

By introducing distributed interaction along with proper social protocols and system policies, its potential for improved collaboration, understanding and discussion could lead to safer and faster medical decisions, with a significant impact for modern healthcare.

This study gathered insights during a controlled simulated team meeting in preparation for follow-up studies during real Pre-operative meetings. It allowed us to collect valuable unbiased, undisturbed and undistracted feedback through a series of individual sessions. Based on the results from this work, we are now refining the techniques and the setup, and plan to run experiments with the next generation of our prototypes in collaborative multi-user meetings.

References

1. Bier, E.A., Freeman, S.: MMM: a user interface architecture for shared editors on a single screen. In: Proc. UIST 1991, pp. 79–86 (1991)
2. Eng, J., Leal, J. P., Shu, W., Yang, G.L.: Collaboration System for Radiology Workstations. *Radiographics* 22, e5 (September 2002); Published online August 21
3. Flanders, A.E., Wiggins III, R.H., Gozum, M.E.: Handheld Computers in Radiology. *Radiographics* 23, 1035–1047 (2003)
4. Frykholm, O., Lantz, A., Groth, K., Walldius, Å.: Medicine Meets Engineering in Cooperative Design of Collaborative Decision-supportive System. In: Proc. CBMS 2010 (2010)

5. Groth, K., Frykholm, O.: Efficiency in Treatment Discussions: A Field Study of Time Related Aspects in Multi-Disciplinary Team Meetings. In: Proc. CBMS 2009, pp. 1–8 (2009)
6. Groth, K., Olin, K., Gran, O., Permert, J.: The role of technology in video-mediated consensus meetings. *Journal of Telemedicine and e-Health* 14(7), 349–353 (2008)
7. Kane, B., Luz, S.: Multidisciplinary Medical Team Meetings: An Analysis of Collaborative Working with Special Attention to Timing and Teleconferencing. *Journal of CSCW* 15, 501–535 (2006)
8. Kane, B., Luz, S.: Achieving Diagnoses by Consensus. *Journal of CSCW* 18, 357–392 (2009)
9. Kane, B., Luz, S., O'Brian, D.S., McDermott, R.: Multidisciplinary team meetings and their impact on workflow in radiology and pathology departments. *BMC Medicine* 5, 15 (2007)
10. Lee, S.-K., Peng, C.-H., Wen, C.-H., Huang, S.-K., Jiang, W.-Z.: Consulting with Radiologists outside the Hospital by Using Java. *Radiographics* 19, 1069–1075 (1999)
11. Li, J., Mansfield, T., Hansen, S.: Supporting Enhanced Collaboration in Distributed Multidisciplinary Care Team Meetings. In: Proc. CBMS 2008, pp. 482–487 (2008)
12. Morris, M. R., Huang, A., Paepcke, A., Winograd, T.: Cooperative gestures: multi-user gestural interactions for co-located groupware. In: Proc. CHI 2006, pp. 1201–1210 (2006)
13. Myers, B.A., Stiel, H., Gargiulo, R.: Collaboration using multiple PDAs connected to a PC. In: Proc. CSCW 1998, pp. 285–294 (1998)
14. Myers, B.A.: Using Hand-Held Devices and PCs Together. *Communications of the ACM* 44(11), 34–41 (2001)
15. Måseide, P.: The deep play of medicine: Discursive and collaborative processing of evidence in medical problem solving. *Communication & Medicine* 3(1), 43–54 (2006)
16. Paek, T., Agrawala, M., Basu, S., Drucker, S., Kristjansson, T., Logan, R., Toyama, K., Wilson, A.: Toward universal mobile interaction for shared displays. In: Proc. CSCW 2004, pp. 266–269 (2004)
17. Olsen, D.R., Nielsen, T.: Laser pointer interaction. In: Proc. CHI 2001, pp. 17–22 (2001)
18. Ratib, O., Michael McCoy, J., Ric McGill, D., Li, M., Brown, A.: Use of Personal Digital Assistants for Retrieval of Medical Images and Data on High-Resolution Flat Panel Displays. *Radiographics* 23, 267–272 (2003)
19. Rekimoto, J.: A multiple device approach for supporting whiteboard-based interactions. In: Proc. CHI 1998, pp. 344–351 (1998)
20. Ruhstaller, T., Roe, H., Thürlimann, B., Nicoll, J.J.: The multidisciplinary meeting: An indispensable aid to communication between different specialities. *European Journal of Cancer* 42(15), 2459–2462 (2006)
21. Stewart, J., Bederson, B.B., Druin, A.: Single display groupware: a model for co-present collaboration. In: Proc. CHI 1999, pp. 286–293 (1999)
22. Tse, E., Histon, J., Scott, S.D., Greenberg, S.: Avoiding interference: how people use spatial separation and partitioning in SDG workspaces. In: Proc. CSCW 2004, pp. 252–261 (2004)
23. Wallace, J.R., Scott, S.D., Stutz, T., Enns, T., Inkpen, K.: Investigating teamwork and taskwork in single- and multi-display groupware systems. *Personal and Ubiquitous Computing* 13(8), 569–581 (2009)

How Technology Influences the Therapeutic Process: A Comparative Field Evaluation of Augmented Reality and In Vivo Exposure Therapy for Phobia of Small Animals

Maja Wrzesien¹, Jean-Marie Burkhardt², Mariano Alcañiz^{1,3}, and Cristina Botella^{3,4}

¹ Instituto Interuniversitario de Investigación en Bioingeniería y Tecnología Orientada al Ser Humano, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain

² Paris Descartes University, LATI, 45 rue des Saints-Pères 75270 Paris cedex 06, France

³ CIBER, Fisiopatología Obesidad y Nutrición, CB06/03 Instituto de Salud Carlos III, Spain

⁴ Departamento de Psicología Basica y Psicobiología Universidad Jaume I, Castellón, Spain

Abstract. In Vivo Exposure Therapy (IVET) has been a recommended protocol for the treatment of specific phobias. More recently, several studies have suggested that Augmented Reality Exposure Therapy (ARET) is a potentially effective technology in this field. The objective of this paper is to report the preliminary results of a comparative analysis of ARET and IVET applied to the treatment of phobia to small animals. To analyze participants' activity, we have adopted a multidisciplinary and mixed perspective based on clinical and user-centered approaches. This pilot results show that ARET and IVET are both clinically effective. Both therapies produce a significant reduction in the clinical outcome measures and allow the clients to interact with a real phobic stimulus after the therapeutic session. The results also show some main differences between technology-mediated therapy and traditional non-mediated therapy. We discuss these results in terms of future design and evaluation guidelines for Mental Health technologies.

Keywords: Mental health, augmented reality, field evaluation.

1 Introduction

It is increasingly recognized that innovative technologies have strong potential in the Mental Health (MH) field [1]. New technologies such as Virtual Reality (VR) or Augmented Reality (AR) can provide therapists with a wide range of MH services and functions to support their therapy and assessment activities. Also, VR and AR allow clients¹ to have easier access to MH services and enhance their engagement in the treatments [2] due to the strong representational and immersion capability of these technologies. However, the way a specific MH system is designed may induce use-related problems. For example, a weak design of Human-Computer Interactions (HCI) may lead to a significant increase in the therapists' or clients' workload during

¹ The word client is usually used in Mental Health Care to describe a person suffering from mental illness.

the clinical intervention. The task of using the system itself may also distract the clients as well as the therapist from the therapeutic session. All of these aspects have stimulated researchers to place more importance on the design guidelines and evaluation techniques for MH technologies (e.g. [1], [2], [3]).

A critical but still rather unexplored issue consists in how these technologies actually support their intended users in their respective (but closely related) activities; how they are actually used; and how they modify the therapeutic process. Among studies of mental health VR/AR systems, most have concentrated on a “client perspective” (e.g. [4], [5]) with a focus on measuring client outcomes as a function of the VR/AR systems and/or therapy orientations. On the other hand, of the few studies in HCI field regarding mental health VR systems (e.g. [6], [7]), most have concentrated on a “therapist perspective” with a focus on measuring or observing the activity of the therapists with the system. Both perspectives are limited in the sense that they consider only a single user at a time, instead of looking at both the therapist and the client working together in a virtual and/or real environment. Therefore, there is a need to propose the “mixed perspective” that should take into account the two actors of the therapeutic process (client and therapist) as well as the technology and the environment with virtual and/or real artifacts. Moreover, to the authors’ knowledge, no studies regarding HCI issues have yet been published in the field of mental health AR technologies.

The aim of this paper is to present a multidisciplinary (i.e., HCI and clinical) and mixed approach (i.e., one that takes into account the two actors of the therapeutic process, the technology, and the collaborative processes that take place among them in a virtual and/or real environment) in a case study of MH technology evaluation. More specifically, this paper reports pilot data on a clinical setting evaluation of Augmented Reality Exposure Therapy (ARET) and In Vivo Exposure Therapy (IVET) for small animal phobias (spiders and cockroaches). The paper is organized as follows. Section 2 presents some main issues regarding In Vivo, Virtual Reality, and Augmented Reality Exposure Therapies applied to the treatment of small animal phobias. Section 3 presents our framework to address the evaluation. Section 4 presents the methodology undertaken in this evaluation. The remainder of the sections present the results and conclusions with a number of design implications for the MH technologies that are drawn from the analysis.

2 In Vivo, Virtual Reality, and Augmented Reality Exposure Therapy in the Treatment of Specific Phobias of Small Animals

Specific phobias all share a common pattern of “marked and persistent fear of clearly discernible circumscribed objects or situations” [8]. Reports on Mental Health present specific phobias as one of the most common single mental disorders [9]. In Vivo Exposure Therapy (IVET) is considered to be standard therapy for the treatment of specific phobias [10]. During this type of therapy, the clients are exposed to real (live) phobic objects or situations, they confront real spiders in arachnophobia or cockroaches in phobia to cockroaches. The effectiveness of IVET has been demonstrated by numerous researchers; however, this treatment has some drawbacks. First, the therapist is not in full control of the real phobic object or the situation. Thus,

it is difficult for the client to have a personalized, graduated exposure in accordance with specific fears (i.e., the only way to present cockroaches or spiders without any movement is to kill them). Second, arranging and organizing the exposure can be time-consuming and can create logistics problems (e.g. feeding the animals; finding a place for the terrarium; etc.). Third, the treatment is associated with a high dropout rate and low level of acceptance [11].

Virtual Reality Exposure Therapy (VRET) might be an interesting option to consider when trying to motivate phobia sufferers to treat their disorders. Indeed, besides having the same effectiveness as traditional therapy [12], VRET resolves some of the problems associated to IVET [4]. First of all, VRET allows therapists to precisely control the phobic stimulus (i.e., the controlled virtual cockroach or spider can stay immobile, can be moved in different directions on a smaller or wider scale, can change size, and can be multiplied as many times as the client and therapist desire). Second, the virtual animals do not require anything to keep them alive; a simple click of the computer button is enough to make them appear. Third, VRET can recreate environments and situations that would be difficult to arrange in traditional treatments such as IVET.

Augmented Reality Exposure Therapy (ARET) has the same advantages as VRET, but it also has some additional characteristics that might be appealing to both clients and therapists. The ARET system allows clients to perceive the real environment and their body with virtual objects (i.e., augmented reality). According to Botella et al. [4], in the case of small animal phobias, this has two great advantages over the VRET applications. First, the real environment does not have to be modeled; therefore, the costs of programming and modeling the application are lower (limiting the modeling and programming to some specific objects such as cockroaches or spiders). Second, perceiving the virtual object in the real environment may have great importance for a better sense of presence and reality judgment, which are recognized as key aspects in this field [13].

ARET has not yet been thoroughly explored, especially regarding HCI issues. Therefore, the aim of this paper is to fill in some of the gaps in this particular research area and to propose a multidisciplinary and mixed perspective approach for evaluating this MH technology under real world conditions.

3 A Multidisciplinary and Mixed Perspective Approach

The goal of the Augmented Reality Exposure Therapy system is to allow both the therapist and the client to collaborate so that the client confronts the phobic stimulus and interacts with it at the lowest possible level of discomfort. In this paper, we are interested in studying two unexplored issues associated to this perspective. First, the design of the system should support the therapeutic relationship between the client and the therapist, which is defined as the therapeutic alliance. In fact, a different interpretation of the technology-mediated therapeutic process is needed. The dynamics of the therapeutic process has until now been analyzed from a purely clinical point of view. This is understandable when the therapeutic process corresponds only to face-to-face traditional client-therapist interactions. However, with the introduction of new technologies to the therapist's office, the dynamics of the

client-therapist interactions may change. Therefore, this new technology-mediated context should be addressed in terms of collaboration. Indeed, the relationship between the client and the therapist, which is defined in terms of the therapeutic alliance concept, is expressed as “(...) the quality and strength of the collaborative relationship between client and therapist (...)” [14]. The quality of this relationship has been shown to contribute to 30% of the positive changes in the client [15]; therefore its quality is of particularly great interest. This point of view is shared by different authors. Doherty et al. [1] proposed using the therapeutic alliance measure between client and therapist in order to gain insight into effects of technology on the therapeutic relationship. Meyerbröker and Emmelkamp [16] analyzed the relationship between the therapeutic alliance and the outcome of the therapy with Virtual Reality Exposure Therapy for specific phobias (i.e., acrophobia and fear of flying).

The second unexplored issue corresponds to the adaptation of the ARET system to face-to-face psychotherapy in an environment mixing both real and virtual objects. Previous HCI research studies in the MH field have not explicitly exploited the “mixed perspective” in order to account for the complex interactions between the therapist, the client, the technology, and the real and/or virtual environments in which they take place. Therefore, a systemic analysis (i.e., mixed) framework is needed in order to fully understand this complex therapeutic context. Recently, Nardi and Kaptelinin [17] showed the benefits of systemic analysis such as Activity Theory in the design of different applications; and Hollan and Hutchins [18] demonstrated the benefits of the Distributed Cognition framework in the design and evaluation of different Augmented Environments. Wrzesien et al. [19] applied a similar approach to analyze Augmented Reality Exposure Therapy. Thus, in our opinion, the application of a distributed theoretical framework has the following advantages. First, it takes into account all possible components of the unit of analysis (i.e., client, therapist, technology, environment, and the interactions among them). Second, the analysis can take place in real world conditions by taking into account day-to-day clinical practice and making the evaluation ecologically valid (i.e. performed in the real-life situation). Finally, the technology is considered to be an equally important component of the unit of analysis, which may positively and/or negatively influence other components.

In summary, the therapeutic process mediated by technology or non-mediated by technology should be analyzed in the following terms. First, the HCI issues and clinical issues should be combined into a multidisciplinary approach. Second, the mixed perspective analysis of interactions between the different poles involved in the therapeutic activity should be proposed. Therefore, the results in the following sections will reflect this double multidisciplinary and mixed approach.

4 Methodology

4.1 Research Design

The study compares two different types of therapeutic processes: the traditional therapeutic process (IVET) and the technology-mediated therapeutic process (ARET). These two therapeutic processes included identical therapeutic objectives and clinical protocols and were located in the same place. In order to reduce the impact of the individual therapeutic style on the development of the client-therapist relationship, all

of the therapists treated clients in both groups. Thus, any differences in clients' clinical outcomes, clients' perceived therapeutic relationship, and therapeutic activities should be attributed to the therapy type factor (technology-mediated vs non-mediated).

4.2 Participants

The participants for this study (clients) were selected according to the DSM-IV [8] criteria for a specific phobia to small animals. In total, 12 clients participated in this study (eleven women and one man, $M=28.54$ years old; $SD=7.92$). Each client was randomly assigned to one of two groups (the ARET group or the IVET group).

Three therapists working in the clinic participated in this study. They all had a minimum of one year of experience in the therapeutic field; however, two of them were novices in the use of the ARET system.

4.3 Materials

All the therapeutic sessions followed the same "one-session treatment" protocol [20]. The protocol involves the use of intensive exposure carried out in one session of a maximum of three hours. The protocol is composed of four parts: (a) exposure; (b) modeling (by demonstrating the interaction with the phobic stimulus by the therapist followed, if possible, by the client); (c) cognitive restructuring; and (d) reinforcement. The main goal of the therapeutic session is to allow the client to confront the phobic stimulus and interact with it at the lowest possible level of discomfort.

a) In Vivo Exposure Therapy. The IVET (see Figure 1a) corresponds to the direct confrontation of a real feared stimulus (i.e., cockroaches or spiders). This type of therapeutic activity involves both the therapist and the client interacting with a real cockroach or spider in order to expose the client to his/her phobic stimulus.

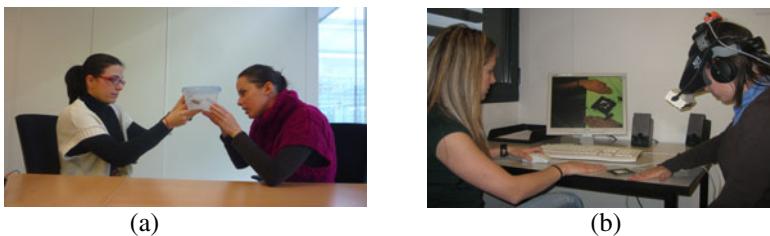


Fig. 1. In Vivo Exposure Therapy (Figure 1a), and Augmented Reality Exposure Therapy (Figure 1b). Therapists (on the left) and clients (on the right) interact with the phobic stimuli.

b) Augmented Reality Exposure Therapy. ARET (see Figure 1b) corresponds to the direct confrontation of a virtual feared stimulus in the real environment (i.e., augmented reality). This type of therapeutic activity involves both the therapist and the client interacting with the virtual cockroach or spider in order to expose the client to his/her phobic stimulus.

The interface that controls the ARET system was only available to the therapist and corresponded to a keyboard. The system allows the therapist to control the phobic stimulus by choosing different functions (on the keyboard) that increase/decrease the number of cockroaches; increase/decrease the size of cockroaches; make the cockroaches move; stop the cockroaches; kill the cockroaches (make the cockroaches look dead); and allow the current action to get back to the beginning. The functions can be combined to adapt the exposure exercises to different conditions according to the client's needs. The client observed the AR environment using a Head Mounted Display (HMD) and actively participated in the exposure exercises proposed by the therapist. However, his/her actions (apart from the viewpoint) did not have any direct influence on the AR environment, which was fully controlled by the therapist. The therapist observed the AR environment (client's viewpoint) on a computer screen. For a detailed technical description of the ARET system, see [4].

4.4 Instruments

The intention of the research team was to collect both qualitative and quantitative data. For the quantitative data, different measures were applied to both the clients and the therapists. The clients' clinical outcome measures corresponded to their anxiety, avoidance of the phobic stimulus, and belief in catastrophic thoughts regarding the phobic stimulus (on the 10-point Likert scale). They also performed a Behavioral Avoidance Test (i.e., BAT, adapted from Öst [20]) in order to define (on a 13-point Likert scale) how close the client was able to approach the real cockroach/spider (e.g. 0/12 corresponds to the client refusing to enter the same room where the phobic stimulus is; and 12/12 corresponds to the client interacting with the phobic stimulus for at least 20 seconds). The clients filled out the short version of the Client Working Alliance Inventory [21] questionnaire, which measures (on a 7-point Likert scale) their relationship with the therapist just after the diagnostic interview (before the therapeutic session) and just after the therapeutic session. The therapists filled out a questionnaire, which evaluates the capacity of the ARET system to help them in constructing a therapeutic relationship with the client (as perceived by the therapist). More specifically, the therapists were asked to rate (on a 5-point Likert scale) the degree to which they agreed with statements that described how the system supports (positive valence) or disturbs (negative valence) the therapeutic alliance between the client and the therapist. This questionnaire was based on the short version of the Therapist Working Alliance Inventory [21]. Moreover, the therapists filled out a questionnaire regarding the usefulness and the frequency of use (on a 5-point Likert scale) of the functions available in the ARET system.

For the qualitative data, the therapists responded to an informal interview related to their experience with the ARET system. A video analysis of therapeutic activity was also applied. In order to preserve the privacy of the participants and not influence the data, all the sessions were recorded using digital cameras and were analyzed (viewed and coded) afterward. No identifiable references to the clients' identities were recorded. More specifically, to record ARET therapeutic activity one video camera recorded a global view and another one focused on the computer screen and the keyboard area. To record IVET therapeutic activity only one video camera was used to record a global view. In addition, commands and interactions of participants with the ARET system

were automatically logged during the session. The video analysis covered approximately twenty-five hours of therapeutic session. The following coding scheme for both verbal and non-verbal behaviour was used to capture the distributed activity between the client, the therapist, and their respective interactions (either mediated by technology or non-mediated by technology). The coded activities corresponded to the following categories: verbal communication, visual attention, and performed actions that involved one or both actors and their environment. Once the videotaped sessions were entirely coded, the frequencies of each event were calculated. The reliability of the coding method was assessed by calculating the correlation among 3 different judges that coded in parallel the same 10-minute extract of the video. The correlation was strong ($R=.930$; $p=.002$ between judges 1 and 2; $R=.966$; $p=.000$ between judges 2 and 3; and $R=.944$; $p=.001$ between judges 3 and 1).

Since the therapeutic sessions varied from 1 hour to 3 hours, all data were presented as frequencies per hour. Because some coded activities were not visible during the whole therapeutic session (the camera angle did not allow showing the entire room), the frequencies of the event were estimated according to the calculated frequencies that could be observed for each session. The logs of the commands and interactions of participants with the ARET system could not be used due to technical problems. Therefore, the analysis of the interactions with the system during ARET was performed by video-analysis, and later confirmed by the respective therapists.

4.5 Procedure

All of the clients participated in the diagnostic interview. Once the diagnostic was confirmed, they were informed about the objectives of the study and had to complete a consent form. The clients also filled out the pre-test questionnaires and performed the BAT, after which they received the therapeutic session (following the one-session treatment protocol) using Augmented Reality Exposure Therapy or In Vivo Exposure Therapy. The session was typically organized as follows. The therapists performed the exposure exercises with the clients. These exercises were previously defined during the diagnostic interview and hierarchically organized from the least anxious for the client to the most anxious. Each exposure exercise was first performed by the therapist (i.e., modeling), then the client was invited to repeat the exercise followed by the cognitive restructuring and reinforcement of the therapist. The client-therapist communication regarding anxiety, irrational thoughts, or other issues took place during the therapeutic session. At the end of the session, the clients filled out the same questionnaires and performed the BAT.

4.6 Data Analysis

The quantitative data corresponded to the clinical measures and the therapists' answers to the questionnaire. To investigate the clinical effectiveness of each therapeutic session, non-parametrical statistic analysis (Wilcoxon) was applied. In order to explore the potential differences between the traditional IVET and ARET groups regarding clinical outcome measures, the differences between pre and post scores were calculated (delta) for each participant, after which the non-parametrical statistic analysis (Mann-Whitney) was used. In order to study the capacity of the

system to help the therapists in constructing a therapeutic relationship with client (as perceived by the therapists), the mean score of four items related to the same dimension was calculated. Afterwards, Cronbach's alpha was applied. Cronbach's alpha was also used to evaluate the agreement of the therapists regarding the usefulness and frequency of use of the functions available in the ARET system. The mean score of four items related to the same dimension was also calculated in order to study the therapeutic relationship between the two actors (as perceived by the clients). Afterwards, Wilcoxon analysis was applied in order to study the evolution of the therapeutic relationship before and after the therapeutic session, and Mean-Whitney analysis was used to compare the therapeutic relationship in two groups. With regard to the qualitative measures, the mean frequencies of events per hour were calculated for each group. The non parametrical statistics (Mann-Whitney) was applied to show significant differences in the therapeutic activity between groups. All analyses were performed using the SPSS 16.0 application with the significance level set at 0.05.

5 Results

5.1 The Nature of Therapeutic Activity

a) Verbal communication. The results in Table 1 show that the verbal communication between the therapist and the client seem to remain similar in both groups ($U=10.00$; $p= .200$ for the therapist verbal communication, and $U=13.00$; $p=.485$ for the client verbal communication). In fact, the therapist-client communication in both groups followed the question-answer model, and the client responded to the therapist's comments.

Table 1. Mean number of verbal communication (VC) per hour and its Mann-Whitney comparison (U) between therapists and clients in each group

	IVET M (SD)	ARET M (SD)	U (p)
Therapists' VC	86,11 (23,30)	119,32 (42,51)	4,00 (0,14)
Clients' VC	79,54 (24,02)	86,33 (25,33)	7,00 (0,46)

b) Visual attention. The video analysis shows that the client's visual attention seemed to be mainly focused on the phobic stimulus in both groups (virtual phobic stimulus vs real phobic stimulus). However, the clients' visual contact with the therapist significantly differed between the two groups ($U=.000$; $p=.004$). During the ARET sessions, the clients spent most of the time looking at the virtual animals without having any, or very little, visual contact with the therapist ($M=0,69$; $SD=0,56$ per hour of session). On the other hand, during the IVET sessions, the clients also spent most of the time looking at the real animals, but they had significantly more frequent visual contact with the therapist ($M=37,75$; $SD=10,81$ per hour of session).

The therapists' visual attention was distributed between more sources than the clients' visual attention. In fact, during the therapeutic session, the therapists distributed their visual attention among different targets. The therapists that participated in ARET seem to look mainly at the following targets: (a) visual sources related to the phobic stimulus (i.e., the computer screen, in order to control the client's view of the AR phobic stimulus; the keyboard, in order to choose the appropriate functions in order to expose the client to the phobic stimulus; and the AR marker on which the virtual animals appeared); (b) notes, in order to write comments and/or clinical measures or keep a temporal record of the client's evolution; and (c) the client, in order to see his/her reactions and possible anxiety. The therapists that participated in IVET seem to look mainly at the following targets: (a) the real phobic stimulus, in order to reference themselves to the object in the conversation, or to control; (b) notes, in order to write comments and/or clinical measures or keep a temporal record of the client's evolution; and (c) the client, in order to see his/her reactions and possible anxiety. As Figure 2 shows, the frequency of the therapists' visual attention was significantly higher for the phobic stimulus target and its related visual sources in ARET, and for the notes in ARET, respectively, $U=.000$; $p=.004$; $U=4,000$; $p=.026$; the therapists' visual attention on client comparison was not significant ($U=8,000$; $p=.132$).

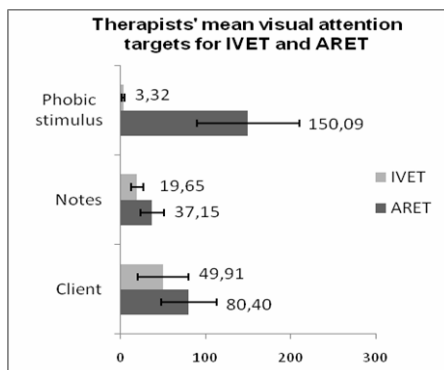


Fig. 2. Mean frequencies of therapists' visual attention per hour in each group

c) Spatial Orientation. Figure 3 (a) and (b) present the spatial orientation of the two actors. As the results show, the therapist and the client seem to move more in IVET than in ARET. Both actors used all the space available in the therapist's office during IVET, while the actors participating in ARET had limited displacements. With respect to the phobic stimulus, the client and the therapist observed the phobic stimulus in two groups from two different perspectives (on the table and on the floor). In the ARET sessions, the clients were able to observe the phobic stimulus on the wall as well.

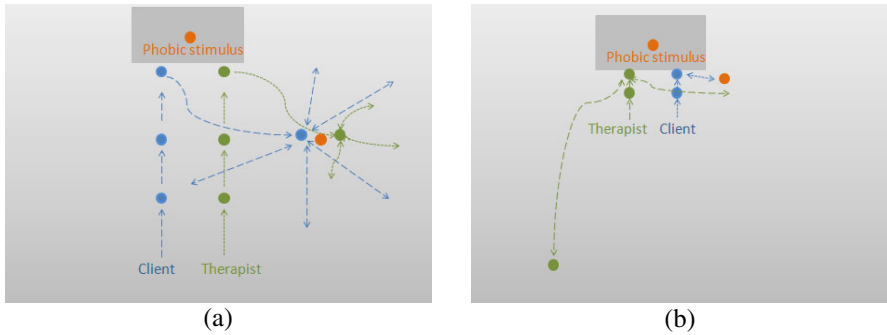


Fig. 3. Example of the spatial displacements of the therapist and client with respect to the phobic stimulus in IVET (Figure 3a) and in ARET (Figure 3b)

d) *Interaction with the phobic stimulus.* Table 2 shows the variation of exposure exercises observed in both types of therapeutic sessions. The results seem to show that, during ARET, therapists proposed more types of exercises than in the case of IVET. The results also show that, during IVET, the therapists proposed to the clients interact with the real phobic stimulus using a piece of paper or a wooden stick. This was not observed in the ARET sessions.

Table 2. Different exposure exercises proposed by the therapists, observed during the IVET and ARET sessions

Exposure exercises	ARET	IVET
Observe immobile phobic stimulus	X	X
Observe mobile phobic stimulus	X	X
Observe dead phobic stimulus	X	-
Observe phobic stimulus on personal belongings	X	-
Put hand(s) near phobic stimulus	X	X
Put foot/feet near phobic stimulus	X	-
Interact with phobic stimulus with an artifact (e.g. a stick)	-	X
Kill phobic stimulus	X	X
Throw away phobic stimulus	X	X
Find phobic stimulus under different artifacts	X	-

5.2 Clinical Effectiveness

a) *Anxiety, avoidance, belief in catastrophic thoughts, and BAT.* The analysis of the pre-test shows no significant differences between the IVET group and the ARET group regarding the anxiety measure ($U=13,000; p=.415$), the avoidance measure ($U=12,500; p=.373$), the belief in catastrophic thoughts measure ($U=10,000; p=.333$), or the BAT measure ($U=9,500; p=.164$). The analysis of the pre-test and the post-test shows that, for the ARET group, all clinical measures decreased significantly after the therapeutic session ($Z=-2.207; p=.027$ for the anxiety measure; $Z=-2.026; p=.026$ for the avoidance measure; and $Z=-2.023; p=.043$ for the belief in catastrophic thoughts

measure). The results also show a significant increase in the BAT scores ($Z=2.232$; $p=.026$) after the treatment (i.e., the clinical improvement on the BAT scale is reflected by an increase of the scores). Also, for the IVET group, the clinical measures decreased significantly after the therapeutic session ($Z=-2.207$; $p=.027$ for the anxiety measure; $Z=-2.014$; $p=.027$ for the avoidance measure). However, for the belief in catastrophic thoughts measure there was no significant difference ($Z=-1.826$; $p=.068$). For the BAT score for the IVET group, the results were significant ($Z=2.232$; $p=.026$).

Figure 4 (a, b, c, and d) shows the comparison analysis of the clinical improvement, which reflects the clinical effectiveness of the two types of therapeutic sessions.

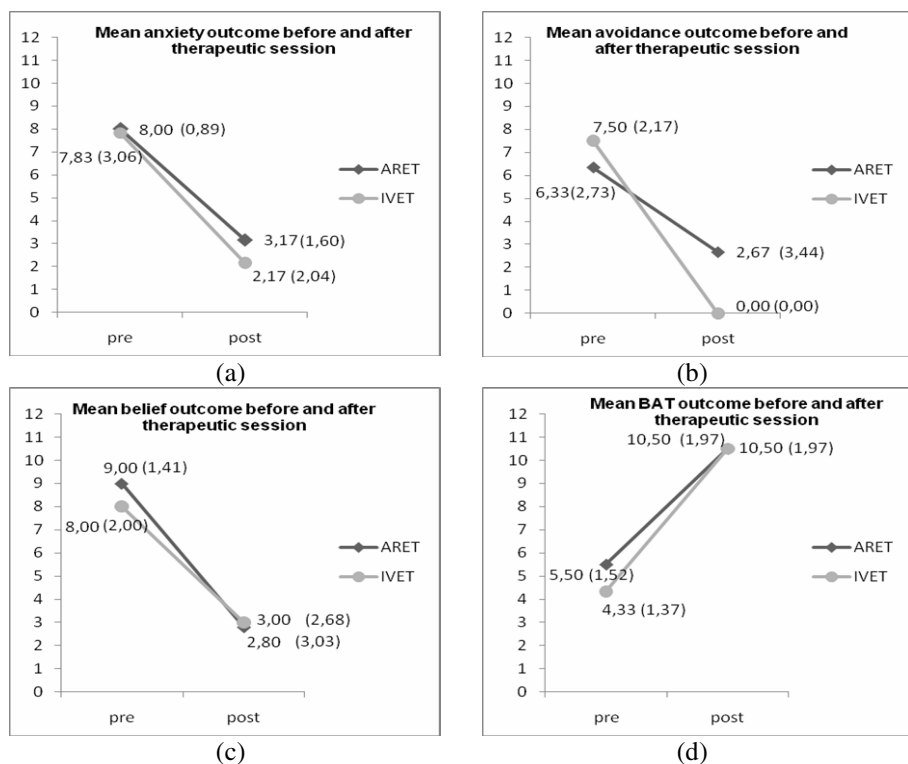


Fig. 4. Mean ratings (SD) for anxiety (Figure 6a), avoidance (Figure 6b), belief (Figure 6c), and BAT (Figure 6d) before and after for both the ARET and IVET sessions

The results show no statistically significant differences for clinical improvement between the two groups for the anxiety measure ($U= 14,500$; $p=.568$), the belief in catastrophic thoughts measure ($U= 13,000$; $p=.792$), and the BAT measure ($U= 8,500$; $p=.115$). However, the clinical improvement of the avoidance score was significantly higher for IVET ($U= 2,000$; $p=.010$) than for ARET.

b) *Therapeutic alliance perceived by the clients.* The analysis of the pre-test show no statistically significant differences between the IVET group and the ARET group with respect to the task measure ($U= 12,000$; $p=.328$), the goal measure ($U= 17,500$; $p=.934$), and the bond measure ($U= 15,000$; $p=.625$). This seems to indicate that, in the two groups (during the diagnostic interview), the therapists and clients created similar high therapeutic relationships.

The comparison analysis of the pre-test and the post-test shows no significant differences in the two groups. More specifically, for the ARET group, all three dimensions of the therapeutic relationship were maintained after the therapeutic session ($Z=0.000$; $p=1.000$ for the task dimension; $Z=-0.405$; $p=.686$ for the goal dimension; and $Z=-0.535$; $p=.593$ for the bond dimension). For the IVET group, all three dimensions of the therapeutic relationship were also maintained after the therapeutic session ($Z=-0.736$; $p=.461$ for the task dimension; $Z=-1.095$; $p=.273$ for the goal dimension; and $Z=-1.656$; $p=.098$ for the bond dimension).

Figure 5 (a, b) shows the comparison analysis of the therapeutic relationship before and after the therapeutic session in terms of tasks, goals, and bond. The results show no statistically significant differences between the two groups neither before the therapeutic session (for tasks ($U= 12,000$; $p=.328$), goals ($U= 17,500$; $p=.934$), and bond ($U= 15,000$; $p=.625$)), nor after (for tasks ($U= 12,500$; $p=.373$), goals ($U= 8,000$; $p=.094$), and bond ($U= 12,000$; $p=.315$)).

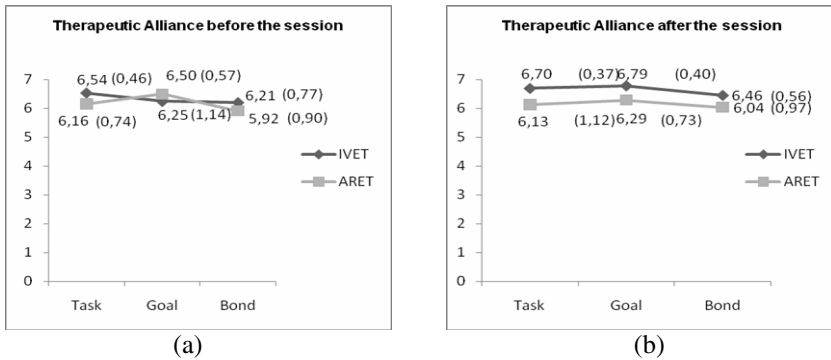


Fig. 5. Mean ratings (SD) for therapeutic relationship in terms of task and goal definitions and bond between the therapist and the client, before (Figure 7a) and after (Figure 7b) the therapeutic session for both the IVET and ARET sessions

5.3 Acceptability

a) *Frequency of use and usefulness of the ARET system.* The results show that the therapists seemed to be in strong agreement ($\alpha=.857$) about using all the functions of the ARET system very frequently ($M=4,88$ out of 5; $SD=0,17$) and seemed to be in less agreement ($\alpha=.573$) about these functions being very useful ($M=4,75$ out of 5; $SD=0,39$). The therapists seemed to be in strong agreement that the function “increase the number of cockroaches by 20” was neither frequently used ($M=1,67$ out of 5; $SD=0,58$) nor useful ($M=2,00$ out of 5; $SD=0,00$). The low score obtained by this function was explained in the informal interview by the fact that the therapists

preferred to use an exposure exercise that was closer to real life (a small number of cockroaches or spiders).

b) The appeal of the ARET system to therapist. The results regarding the appeal questionnaire based on the short version of the Therapist Working Alliance Inventory [21] show how, in the therapists' opinion, the ARET system helps them to create the therapeutic relationship with the client in terms of therapeutic alliance (i.e., task, goal and bond). The general internal consistency of the scores given by the therapists was high ($\alpha=.961$). Therefore, the results seem to show that the therapists agree that the system helps the clients in performing therapeutic tasks ($M=3.92$ out of 5; $SD=1.23$), and strongly helps in defining the goals of the therapeutic session ($M=4.42$ out of 5; $SD=0.80$). The results also seem to show that, in the therapists' opinion, the system moderately helps in creating and maintaining the bond between the client and the therapist ($M=3.42$ out of 5; $SD=1.46$).

c) Usability issues of the ARET system. The informal interviews with the therapists as well as the video analysis brought to light some interesting information about the ARET system. According to the therapists, the ARET system itself and its functions were useful for this type of therapy. More specifically, the therapists were satisfied with the system, particularly regarding the sense of controlled and secure context that the ARET system gave their clients. However, the video analysis showed that the therapists used different strategies while performing the same exposure exercise (using the same functions). After deeper analysis of this issue with therapists, we came to the conclusion that the user interface dialogue of the ARET system was not optimal. More specifically, the functions used by the therapists to present the exposure exercises to the client were not optimal. The first usability issue corresponded to the function "Initiate", which allows the therapist to go back to the beginning of the exercise. This function was also identified as a necessary step before increasing the number of cockroaches/spiders. This had its consequences in terms of cockroach/spider movement. Since the "Initiate" function stops the movement, the therapist had to press the "Move" key one more time in order for the phobic stimulus to regain movement.

The second usability issue corresponded to the function that allows the therapist to kill the cockroaches/spiders ("Kill"). This function was used when the clients were smacking cockroaches/spiders with the swatter in order to kill them. The problem appeared when the therapist asked the client to kill more than two animals (i.e., the function worked for only two animals during the same exposure exercise). Once the therapists realized that the "Kill" function only worked for two cockroaches/spiders, they had to find a different strategy to reduce the client's frustration and the possible clinical consequences of not being able to kill all of their phobic stimuli (i.e., using the "stop" functions or "initiate" function).

The third usability issue corresponded to the visibility of the dark animal on the black AR marker. Two techniques were identified as a solution for this concern. The first technique was to make the animal larger by using the "Increase the size" function in order to see the cockroach/spider outside of the black part of the AR marker. The second technique consisted of moving a cockroach/spider a few centimeters by pressing the "Move" function and then stopping it outside of the AR marker by

pressing the “Stop” function. The therapists found these usability issues confusing, making the ARET system less intuitive. They also considered them to be time-consuming.

The therapists recommended several improvements for the system, such as more unpredictable trajectories of animals, additional behavior (e.g. flying cockroaches), incorporation of sounds (e.g. during walking, killing), incorporation of different varieties of animals (there are currently only 3 types of spiders, and 1 type of cockroach). Finally, the therapists mentioned the need for lighter and smaller HMD to reduce the clients’ level of tiredness.

6 Discussion and Conclusions

This study evaluates a technology-mediated therapeutic activity (Augmented Reality Exposure Therapy) and a non-mediated by technology therapeutic activity (In Vivo Exposure Therapy) in real-world conditions. The main findings and their implications are discussed below.

The first objectives of this paper were to study how the ARET system supports the therapeutic relationship between the client and the therapist and to determine if the presence of the technology influences the clinical outcome. By applying the multidisciplinary approach, we aimed to demonstrate that, in both technology-mediated therapeutic sessions (ARET) and non-mediated by technology therapeutic sessions (IVET), the relationship between the client and the therapist were the same. More specifically, the results seemed to show that the therapeutic alliance score (evaluated by clients) was similar in both groups after the diagnostic interview (i.e., pre-test measure) and remained similar after the therapeutic session (i.e., post-test measure). Therefore, even though some hesitant clinicians believe that the therapeutic alliance is at risk due to the introduction of technology to the therapeutic process [22], this pilot data is a first step in demonstrating that the therapeutic alliance can be created very well in both treatment conditions and that there are no significant differences between them. The clinical outcome seems to confirm these results. Indeed, the clinical symptoms (i.e., anxiety and avoidance) significantly decreased after the therapeutic session, and the BAT score significantly increased in both groups. Moreover, there was no significant difference between the two groups. The only differences were observed in the avoidance score, which decreased significantly more in the IVET group than in the ARET group. This result might be due to the small sample size of the clinical population used in this study. This assumption can be explained by the fact that these results can be contrasted with direct confrontation to the real phobic stimulus (BAT score). The BAT score results demonstrate that, in both groups, the clients did not avoid the real phobic stimulus after the therapeutic session, but rather interacted with it. In our opinion, the lack of significant differences for the belief in catastrophic thoughts measure in the IVET group is also related to the small size of the clinical sample.

Although some usability issues of the ARET system were detected (i.e., non-optimal user-interface dialogue), overall the therapists evaluated the system as being particularly useful and acceptable. More specifically, the results seem to show that the ARET system provides strong insight into the client-therapist relationship in terms of

task and goal definition. In our opinion, the main reason for this conclusion is the flexibility of the system, which proposes various exposure exercises to be presented from different perspectives (on the floor, on the table, and on the wall) to each client, and which allows tasks and goals to be followed. We hope that with the introduction of the changes required by the therapists and improvements in the user-interface dialogue, the system will also provide strong insight into the client-therapist relationship in terms of bond.

The second objective of this paper was to study how the technology-mediated therapeutic activity differs from non-mediated by technology therapeutic activity. The study shows the importance of the distributed, mixed perspective approach in the analysis of this specific collaborative work activity. Indeed, addressing these issues requires observing the activity of the actors in the context of real therapeutic situations. The usual HCI methods that are used to evaluate VR and/or AR applications (e.g. inspections, verification of design guidelines, testbed evaluations) are limited in MH applications in the following way. First, the MH technology evaluation should include day-to-day therapeutic practice. This involves the participation of the clinical population, which is fragile and heterogeneous. Second, the objective of the evaluation should not only consider performances related to the specific sub-tasks (i.e., clinical efficacy), but a larger context related to the client-therapist relationship. Finally, the evaluation involving the clinical population has some clinical and/or ethical constraints that should also be taken into account.

The results show several interesting differences between the ARET and IVET sessions that can be interpreted in terms of future MH technology design guidelines. First, the verbal communications between the clients and the therapists seemed to be the basis of the therapeutic session in both groups. This is understandable since both cognitive reconstruction and reinforcement are one of the most important parts of the phobia treatment. Second, the visual attention of the therapists seemed to differ from the clients' visual attention in the same way in both groups. The therapists frequently changed their visual source of information while the clients focused mainly on one visual source (phobic stimulus). This difference is understandable since the role of both actors in the therapeutic process is different. However, we noticed that the therapists from ARET had more visual attention targets regarding the phobic stimulus (i.e., computer screen, keyboard, and AR marker on which the VR animal appears), and spent significantly more visual attention on it than the therapists from IVET (i.e., real phobic stimulus). Since frequent and numerous switches between different tasks makes the process more demanding for the user and might be a sign of over-load [23], reducing the sources in ARET should be considered in order to make more efficient use of the therapists' resources. Moreover, the clients from the ARET group seemed to have significantly lower visual contact with the therapist than in the IVET group. This result might be explained by the fact that the clients in the ARET group wore HMD, which limited their visual field and head movements. However, even though visual contact plays an important role in all types of face-to-face communication, and its use demonstrates engagement as well as attention and liking (e.g. [24]). The limited amount of visual contact does not seem to influence the clinical effectiveness of the ARET or the client-therapist relationship. The results also seem to show that ARET differs from IVET in terms of spatial displacement. Indeed, the limited spatial movements in the ARET group can be due to the numerous cables and movement

limitations that the system places on the user. However, the displacements in IVET were not only due to fewer constraints on the client, but also due to the therapist trying to control the live phobic stimulus and the client getting more anxious. This might explain why the displacements were so numerous. Finally, the activity analysis seems to show that the ARET system allows therapists to propose a wide range of varied exposure exercises, which in turn allows the client to be exposed to numerous real-life situations that would be difficult to propose with a real, less controllable phobic stimulus. This is an important advantage that can still be improved by the introduction of more direct interactions with the virtual phobic stimulus, such as those present in IVET sessions (i.e., direct interaction with a wooden stick or a piece of paper). Direct interactions of this type can also be developed in terms of simple exposure therapy games, during which the client can interact with the phobic stimulus in a more fun context. In fact, in their recent pilot clinical study, Botella et al. [25] showed the positive effect of a mobile therapeutic game on the client's clinical measures in a case study of cockroach phobia treatment. Thus, the enjoyment factor related to the exposure seems beneficial for clients and can open up a new range of related research.

The results show that applying both a multidisciplinary and mixed perspective approach can bring to light a great amount of interesting information for future design. In fact, this clinical setting evaluation has identified several shortcomings that highlight the need to improve both the design and the evaluation methods of the MH technologies. First, there is a need to maximize non-clinical, usability evaluation. Indeed, even though the ARET system involved a collaborative design process in which both MH professionals and designers were involved, some usability issues were only detected once the finished ARET system was tested in a clinical setting. Thus, an evaluation using scenario-based tests to detect all possible interaction issues should be performed to help improve requirements gathering. Second, there is a need to take into account all the actors participating in the therapeutic process. While the therapist plays a very important role in the design and evaluation process [1], the therapeutic activity should be analyzed by taking into account both the client and the therapist and their respective and related interactions. More specifically, the interactions between these two actors should first be understood in the traditional clinical setting, and then translated to the technology-mediated setting in order to support these interactions. Similarly, in our opinion, the comparison evaluation such as the one presented in this study is beneficial for future design recommendations. Third, the introduction of more natural user interfaces and the development of innovative interaction metaphors that allow more natural and direct client-therapist interactions (i.e., without HMD, cables, computers screen) should be applied. According to Riva et al. [27] in order to allow the technology-mediated therapeutic process to be effective, the mediation of the technology should disappear from the client's awareness (*disappearance of mediation*). When this phenomenon occurs, the user is not simply observing the projected virtual environment but is actively participating in the therapy. The therapist-client interactions are one of the main factors in creating this phenomenon. Also, therapist-client interactions create *common ground* between the two actors, which is so important to clinical experience [26]. Finally, the design of the system should take into account a large clinical population. Even though the ARET system is presented in a case study for cockroach and spider phobia, its advantage is based on the fact that the simple introduction of different

virtual objects such as snakes, bats, or rats would allow the system to treat different small animal phobias. This flexibility should be considered in all MH applications.

This study can be improved in several ways. First, all the conclusions must be confirmed with a larger clinical population sample. Second, in addition to the frequency of events, additional measures such as time would be interesting to consider. Finally, the appeal of the ARET system to the clients in term of therapeutic alliance could also be evaluated. Even though this preliminary study has limitations, the evaluation process gave us a lot of interesting information in terms of design and evaluation guidelines and showed the importance of taking into account the multidisciplinary and mixed perspective approach in new MH technologies evaluation.

Acknowledgements. We would like to thank to all therapists and clients that participates in this study. We would also like to express a special gratitude to Juana Maria Breton Lopez and Patricia Mesa Gresa for their valuable help. This study was funded by Ministerio de Educación y Ciencia Spain, Project Game Teen (TIN2010-20187) and partially by projects Consolider-C (SEJ2006-14301/PSIC), “CIBER of Physiopathology of Obesity and Nutrition, an initiative of ISCIII” and Excellence Research Program PROMETEO (Generalitat Valenciana. Conselleria de Educación, 2008-157).

References

1. Doherty, G., Coyle, D., Matthews, M.: Design and evaluation guidelines for mental health technologies. *Interacting with Computers* 22(4), 243–252 (2010)
2. Coyle, D., Doherty, G., Sharry, J., Matthews, M.: Computers in Talk-Based Mental Health Interventions. *Interacting with Computers* 19(4), 429–586 (2007)
3. Coyle, D., Doherty, G.: Clinical evaluation and collaborative design: developing new technologies for mental healthcare interventions. In: *Proc. CHI*, pp. 2051–2060 (2009)
4. Botella, C., Juan, M., et al.: Mixing Realities? An application of Augmented Reality for the treatment of cockroach phobia. *CyberPsychology & Behavior* 8(2), 161–171 (2005)
5. Botella, C., Bretón-López, J.M., Quero, S., Baños, R.M., García-Palacios, A.: Treating Cockroach Phobia With Augmented Reality. *Behavior Therapy* 41(3), 401–413 (2010)
6. Brinkman, W.P., Sandino, G., van der Mast, C.: Filed observations of therapists conducting virtual reality exposure treatment for the fear of flying. In: *Proc. ECCE* (2009)
7. Paping, C., Brinkman, W.P., van der Mast, C.: An Explorative Study into a Tele-delivered Multi-patient Virtual Reality Exposure Therapy System. In: *Wounds of War II*, pp. 203–219. IOS press, Amsterdam (2010)
8. American Psychiatric Association: *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Text revision. American Psychiatric Association, Washington, D.C (2000)
9. Alonso, J., Angermeyer, M.C., Bernert, S., et al.: Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. *Acta Psychiatrica Scandinavica* 109(420), 21–27 (2004)
10. Van Hout W.J.P.J., Emmelkamp P.M.G.: Exposure in Vivo Therapy. *Encyclopedia of Psychotherapy*, 761–768 (2003)

11. Choy, Y., Fyer, A.J., Lipsitz, J.D.: Treatment of specific phobia in adults. A comprehensive review on the treatment of specific phobia. *Clin. Psychol. Rev.* 27, 266–286 (2007)
12. Powers, M.B., Emmelkamp, P.M.G.: Virtual reality exposure therapy for anxiety disorders: A meta-analysis. *Journal of Anxiety Disorders* 22(3), 561–569 (2008)
13. Baños, R.M., Botella, C., Garcia-Palacios, A., et al.: Presence and reality judgment in virtual environments: a unitary construct? *CyberPsychology & Behavior* 3(3), 327–335 (2000)
14. Horvath, A.O.: The Alliance. *Psychotherapy* 38(4), 365–372 (2001)
15. Assay, T.P., Lambert, M.J.: The empirical case for common factors in therapy: Quantitative Founding. In: Duncan, B.L., Hubble, M.L., Miller, S.D. (eds.) *The Heart and Soul of Change*, pp. 23–55. American Psychology Association, Washington, DC (1999)
16. Meyerbröker, K., Emmelkamp, P.M.G.: Therapeutic Process in Virtual Reality Exposure Therapy: The Role of Cognitions and The Therapeutic Alliance. *Journal of Cybertherapy & Rehabilitation* 1(3), 247–257 (2008)
17. Nardi, B., Kaptelinin, V.: *Acting with Technology: Activity Theory and Interaction Design*. MIT Press, Cambridge (2006)
18. Hollan, J.D., Hutchins, E.L.: Opportunities and Challenges for Augmented Environments: A Distributed Cognition Perspective. In: Lahlou, S. (ed.) *User Friendly Environments: From Meeting Rooms to Digital Collaborative Spaces*. Springer, Heidelberg (2009)
19. Wrzesien, M., Burkhardt, J.M., et al.: Analysis of Distributed-Collaborative Activity during AR Exposure Therapy for Cockroach Phobia. In: *Proc. Cyberpsychology* (2010)
20. Öst, L.G.: Rapid treatment of specific phobias. In: Davey, G.C.L. (ed.) *Phobias: A Handbook of Theory, Research, and Treatment*, pp. 227–247. Wiley, New York (2000)
21. Tracey, T.J., Kokotovic, A.M.: Factor structure of the Working Alliance Inventory. *Psychological Assessment* 1, 207–210 (1989)
22. Germain, V., Marchand, A., Bouchard, S., Guay, S., Drouin, M.S.: Assessment of the Therapeutic Alliance in Face-to-face or Videoconference Treatment for Posttraumatic Stress Disorder. *Cyberpsychology, Behavior, and Social Networking* 13(1), 29–35 (2010)
23. Neerincx, M.A., van Besouw, N.J.P.: Cognitive task load: a function of time occupied, level of information processing and task-set switches. *Engineering Psychology and Cognitive Ergonomics* 31(6), 247–254 (2001)
24. Piper, A.M., Hollan, J.: Analyzing Multimodal Communication around a Shared Tabletop Display. In: *Proceedings of ECSCW*, pp. 283–302 (2009)
25. Botella, C., Breton-López, J., Quero, S., Baños, R.M., Garcia-Palacios, A., et al.: Treating cockroach phobia using a serious game on a mobile phone and augmented reality exposure: A single case study. *Computers in Human Behavior* 27(1), 217–227 (2011)
26. Riva, G., Zurloni, V., Anolli, L.: Client-Therapist Communication in Computer assisted environment. In: Anolli, L., et al. (eds.) *The Hidden Structure of Interaction: From Neurons to Culture Patterns*. IOS Press, Amsterdam (2005)

You've Covered: Designing for In-shift Handoffs in Medical Practice

Yunan Chen^{1,2}

¹ Department of Informatics

² Institute for Clinical and Translational Science

University of California, Irvine

Irvine, CA 92697-3440, USA

yunanc@ics.uci.edu

Abstract. Handoffs are moments of critical transition in which clinicians engage to maintain continuous coverage of patient care. This paper reports on an observational study of continuous coverage in an Emergency Department (ED), where three types of handoffs that occur during the same shift were identified: lunch breaks, ad hoc breaks and high workloads. The findings show these “in-shift handoffs” are managed not only through temporal linear coordination, but also through the local coordination among nurses working nearby. In-shift handoffs are crucial to maintaining continuous coverage in hospital settings. However, insufficient understanding of in-shift handoffs in Electronic Medical System (EMR) design may lead to a separation of information and responsibility, and an illusion of communication in patient care. The findings of this study call for attention to in-shift handoffs in future system design and for improving the traditional handoff process through the coordination of local awareness during ED work.

Keywords: In-shift Handoffs, Electronic Medical Record (EMR), Emergency Departments, Non-working Moments, Design.

1 Introduction

One unique practice that distinguishes hospital work from other domains of health care is the notion of “continuous coverage” [1]. In his work regarding the temporal organization of medical care, Zerubavel writes, “*since the hospital’s raison d’être is patient care, and since patients require that care regardless of the time of the day, the day of the week, or the time of the year (that is, even at night, on weekends, and on holidays), hospitals must always be open and provide medical and nursing coverage on a continuous basis.* [1: P40].” Indeed, hospital service can never stop, yet no employee can work continuously without breaks. Continuous coverage in a hospital is maintained through the rotation of working/non-working times amongst a group of clinicians. In most hospital departments, nurses and doctors rotate their work to provide non-stop service to patients. At the Emergency Department (ED) where this study was carried out, nurses are required to stay at their patients’ bedside constantly in order to monitor potentially unstable situations. In other words, continuous

coverage requires that the end of one shift continue immediately with a new shift, and that an absence of even a few minutes needs to be covered by other nurses.

The notion of continuous coverage raises critical considerations regarding the design of information systems in hospital environments, since the design needs to consider not only the work performed through the system, but also how different non-working activities are handled. The most studied non-working activity is perhaps the end of shifts *handoffs*. Handoff is defined as when “two or more workers exchange mission-specific information, responsibility, and authority for an operation [2].” Handoffs bring interruptions into the continuous care process since the “baton” of patient care can easily be dropped during these moments of transition [3][4].

Shift changes are the most common handoffs in hospital work, in particular, those for nursing shifts in inpatient units e.g. [5][6]. Yet, continuous coverage in medical work may involve other kinds of handoffs depending on the urgency of the medical service. As Zerubavel mentions, “*having direct responsibility for patients, house staff and nurses never leave their service even for a short time period such as a lunch break without having some ‘cover’ for them* [1: P42].” A lack of consideration for these other types of non-working activities in clinical systems design would inaccurately reflect the work practice being conducted in hospitals.

To understand how continuous coverage is maintained in hospital work, we studied work practices in an Emergency Department (ED). ED patients are often in unstable and life-threatening situations, and even a few minutes of delay in giving medications can be crucial and lead to patient sufferings and care deterioration. Thus, the non-work activities in an ED are not only limited to shift changes, but also include various breaks that occur during the same shift of work. In this paper, we refer to the handoffs that happen during the same shifts as **In-shift Handoffs**. We identify three types of in-shift handoffs and discuss how they are coordinated among ED clinicians. These in-shift handoffs, however, have not yet been reported in current HCI/CSCW literatures. The findings of this study provide valuable insights for designing information systems that support work practice in urgent, life-critical, and time-constraint environments and suggest how information systems can be designed to support the working/non-working moments transitions to better facilitate the practice of continuous coverage in hospital work.

2 Related Work

It is commonly believed in HCI field that understanding work practices and user behaviors is key to the design of systems situated in the real working environment. One critical issue in medical work is that it is highly collaborative and often coordinated among multiple clinicians. Hence, researchers in the HCI/CSCW field have long been interested in studying collaboration and coordination in hospital work. Previous studies have suggested that articulating work is essential to successfully ensuring clinical collaboration and coordination [7]. In particular, hospital work is often managed through temporal coordination [8] and spatial coordination [9].

Maintaining continuous coverage is certainly a coordination issues and it is often studied in the context of shift changes handoffs. A handoff process contains the transition of not only information, but also responsibility and authority between two or more people [2]. Handoffs are often considered as “drop points” in continuous patient

care since the transition of care often occurs within short time periods and under pressure [10]. Medical literatures show that handoffs carry a high risk of communication and information breakdowns [11][12]. In particular, one study shows that two-thirds of communication issues in a clinical setting are related to handoffs [13], indicating the importance of handoffs to not only the work continuity, but to the quality of patient care and patient safety. Because of that, how to ensure the continuity of care during handoffs is key to the quality of medical practices, issues such as communication patterns [6][10][14], information flow [5][15], artifacts usage, [5][6] as well as the sense-making process during handoffs [3] are explored in previous literature.

HCI/CSCW field take handoffs, like shift changes, as a temporal rhythm that spins in the process of medical work and leads to information breakdowns in team collaboration [8]. Other HCI studies show that nurses spend almost an entire hour preparing for shift changes. They often deploy electronic-, verbal-, and paper-based artifacts to facilitate the information transition during shift changes [5] [6]. Randell et al. argues that in addition to handoffs, cognitive artifacts, *e.g.*, the doctor's list and the white board, could provide continuous awareness among clinicians and consequently remove the tension away from handoffs transition moments [16]. Using cognitive artifacts to provide continuous awareness suggest a new possibility of how handoffs could be handled beyond guiding the transition moments.

Beyond handoffs, other types of non-working moments may also happen when work efficiency is "*impaired by over work, tiredness, even boredom derived from the work* [17]." Understanding how other non-working moments are coordinated and covered is important for designing information systems to support continuous coverage in hospital work. This important aspect of work practice, however, has not yet been reported in the current HCI literature.

3 Methodology

This study took place at an Emergency Department (ED) in a large regional hospital. The ED service is generally non-stop, serving on average 200 patients every day. This high patient volume helps produce circumstances of continuous service. The ED is equipped with a centralized Electronic Medical Record (EMR) system for the purpose of keeping patient records as well as facilitating clinical work practices.

3.1 Study Site: Main ED Unit

There are 4 major units in the ED: a pediatric ED, an urgent care unit, a trauma center, and a main ED unit. Each unit has 8-16 patient rooms for the types of patients it serves. Beyond the four major patient treatment units, there is also a waiting room, triage rooms, an admitting office, a meeting room and an administrative office in the ED. Due to the size of the entire ED, this study took place primarily in the main ED. Due to the size of the entire ED, this study took place primarily in the main ED area, where most of the attendings, nurses and ED technicians, and other ED staff, such as discharge manager, social worker, nursing manager and admin personnel are located. The main ED has 2 nursing stations and a MD station, surrounded by 16 patient rooms (see Figure 1).

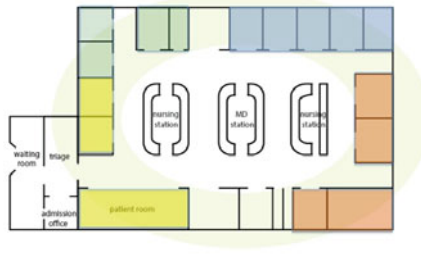


Fig. 1. The main ED map with the work locales of nurses. Four room nurses are each represented in one color and the float nurse is covering for the entire ED.

3.2 Participants

The ED has a total of 57 clinical staff members, with some non-clinical staff. Studies were conducted in two phases. Initially, we observed general ED work in public areas such as the meeting room, the MD station, waiting rooms, nursing stations and hallways. These observations covered most of the ED staff, and they helped us to gain an understanding of overall ED patient care. The second phase consisted of shadowing individual doctors and nurses working in the main ED area, triage room, and waiting rooms. For the research reported in this study, we shadowed 5 attending physicians, 5 room nurses, 4 triage nurses, and 2 float nurses. Other ED staff such as admission personnel, social workers, and the discharge manager were also observed, but are not reported due to the study's emphasis on the continuous coverage of patient care.

3.3 Data and Data Collection

A total of 120 hours of observations were conducted over a period of eight weeks, with 40 observation-hours dedicated to understanding how patient care is performed in the ED in general, and 80 hours of shadowing individual ED staff. Observations were divided into 4-5 hours long sessions where each session focused on one ED staff or location. During observations, we tracked down important incidents happening in the patient care process and asked ED staff to explain their work practices when patients were not around. We also attended daily nursing shift meetings. Observation notes were jotted down using paper and pen on site, and transcribed in detail later.

To ensure accurate representation, approximately half of the observations were carried out during ED peak times – weekends and Monday nights. In addition, we shadowed all roles involved in direct patient care, *e.g.* triage nurses, room nurses, ED doctors and waiting room nurses, as well as, all major locations on the ED floor, *e.g.* patients' waiting rooms, triages, nursing stations, MD stations and patient rooms. To protect patients' confidentiality, we stayed outside patient rooms when shadowing room nurses. However, since the EMR workstations were mounted at the corner of the patients' room facing outside, communication and documentation patterns could easily be observed without interfering with direct patient care.

3.4 Data Analysis

After continuous coverage was identified as the main theme of this paper, instances that related to covering and handoffs activities were extracted from the observation notes. These activities were further categorized into three types according to their length and properties. It is notable that the findings in this study are not drawn from one individual ED staff; rather, they are based on a synthesis of the recurring behaviors demonstrated by multiple staff and repeated observations.

4 Findings

This section first introduces the key staff involved in the process of handoffs, and how their intertwined roles result in “continuous coverage;” then we describe shift handoffs as conducted through the EMR system and outline three types of inshift handoffs and articulate how they are coordinated on the ED floor.

4.1 Patient Care in ED

The continuous coverage of patient care starts as soon as patients arrive in the ED, where they are monitored in the waiting room and the triage rooms. Once patients are assigned a patient room, they are treated by a room nurse and an ED doctor. We use the term **work locale** to represent the physical coverage of ED staff's service area.

ED Doctors. The ED doctor's work largely consists of what Strauss calls “therapeutic work” [17]. ED doctors choose patients based on medical urgency, patient volume and their own expertise. Patients waiting to be seen by the doctor all reside in ED patient rooms. Each doctor, on average, handles 6-7 patients during the ED peak time and fewer patients during non-busy hours. After taking on a new patient, the ED doctor first travels to the patient room to do initial diagnosis work, e.g. interviews, physical exams, and then returns to the MD station to place orders in the EMR system, such as medications, procedures and labs. Orders will be sent to room nurses as pending tasks in the EMR. Doctors spend most of their time checking nursing notes, waiting for patient lab test results, discussing cases and writing progress notes at the MD station.

The ED doctors' work locale extends to the entire ED floor, since they do not pick up patients according to their locations. Doctors may travel to the furthest patient room or take care of a patient right across the MD station. Their movements in the ED are largely centered around the MD station, although they also work on all four ED units. In other words, a doctor's work does not anchor them at bedside at all times. When urgent situations occur, doctors can be paged through either a personal page or an overhead page announced to the entire building.

Room nurses. Each room nurse manages four patients' rooms in the ED. Nurses largely engage in the so-called “monitoring work” [17]. The reason we refer to room nurses' tasks as “monitoring work” is that they station themselves at patient bedside, continually monitoring patients' status, such as watching out for patient situations,

checking vital signs every two hours, updating patient information on the nursing notes section of the EMR, and administrating orders received from ED doctors.

Room nurses, as observed in this study, are always checking the EMR system to see if they have any pending orders and to update nursing notes after every procedure they perform, whether at the bedside or at the nursing station. Each room nurse's service area is marked on Figure 1 using different colors. As indicated on the map, patient rooms assigned to each nurse are all located at the same corner of the main ED. Room nurses divide their entire working hours among patient rooms and the nursing station near their assignments. The typical routine of the room nurse is to check patients, update information at the bedside computer, follow administrative orders, and then write notes at the nursing stations.

Unlike the ED doctors who work on the entire ED floor, the work of the room nurses remain situated around the area of their assignment. This is primarily due to the nature of the monitoring work that must be done at the patients' bedside. For instance, the four room nurses in the main ED area are each in charge of four patients' rooms and there is no overlap among their room assignments (see Figure 1). In other words, at any given moment, there is only supposed to be one room nurse for each patient room. Thus, the working locale for a room nurse is the fan-shaped area from the nursing station to their patients' rooms. Nurses rarely step out of their designated working locales because they need to be readily available to monitor patient situations and to respond to patients' requests. This continuous coverage occurs even during the non-peak time, since nurses have to be prepared to help with any sudden emergencies, such as patient transported by ambulances.

Float nurse. The float nurse is an indispensable role in the process of maintaining continuous coverage. During every shift, one nurse is assigned the role of float nurse by the nursing manager. A float nurse receives no particular assignment, but covers for a room nurse whenever they leave their designated working locale during a shift, or when a room nurse needs help with a heavy workload.

A float nurse's working locale includes all the places that need covering, which in most cases, is the entire ED. Differing from doctors who mostly reside in their own station, a float nurse does not have a fixed working station, and is always moving around on the floor during a shift. Float nurses may actively approach room nurses to see if they want to be covered, or may wait to be called on for help.

4.2 Continuous Coverage in the ED

In the ED, a patient has to be continuously covered on both the therapeutic level and the monitoring level. Before the introduction of the EMR, doctors and nurses had to constantly communicate. Nowadays, doctors and nurses do not have to verbally update each other, since all work is communicated through the system. When an ED doctor signs an order for a patient in the EMR, the room nurse for this patient is automatically alerted. Similarly, when the room nurse of this patient updates notes in the system, the doctor can see them immediately at the MD station. Sometimes, doctors may not even notice who the room nurse of a particular patient is, since their working locale can be across the entire ED, and every patient of theirs may have a different room nurse.

ED doctors are considered on duty as long as they stay in the hospital area, and have no need to be covered by others. In contrast, room nurses have to stay close to patient rooms for their entire service period in order to monitor patients (with the exception of the float nurse). Each work area only has one nurse, so whenever a nurse steps out to get water, take a brief break or have lunch, a disruption in the continuous coverage of patient care occurs; to avoid this, the room nurse needs to be covered before they can leave. In other words, an in-shift handoff is needed.

What follows are our findings regarding handoffs. First, we describe end-of-shift handoffs using the EMR system. Second, we detail three types of in-shift handoffs and their non-EMR-oriented coordination processes.

4.3 End of Shift Handoffs through EMR

Unlike nursing shift meetings in in-patient units that take more than 30 minutes [5] [6], the shift meetings in the ED usually last only 2-3 minutes and are held in the locker room next to the main ED area. Attendees of these meetings are all the incoming shift nurses and the nursing manager on duty. The locker room is where the incoming nurses first arrive when they start their workday. During the meeting, the nursing manager brings in a paper-schedule and announces each incoming nurse's assignment. A typical assignment looks like this¹: Mike - triage room 1, Melissa - float nurse, Rowena - room 12-16, Lisa - room 16-20 and so forth. The nursing manager will then let each nurse know the time of his or her assigned lunchtime – the one-hour middle-of-shift meal. If there are no questions regarding the assignments, the nurses walk to their designated work locales to start their work. The earlier shift nurses in the ED never attend the shift meetings, but wait at their work locale until the incoming shift nurse arrives. Hence, nursing handoffs in the ED are all conducted at the local level where two shift nurses assigned to the same area hand over jobs at the working locale.

Since the deployment of the EMR system, the handoff between nurses now occurs in two stages: an informal verbal information handoff at patient bedside and a formal system handoff at either bedside or the nursing station. During the handoffs, the leaving nurse will walk through each patient room with the incoming nurse. For each room, the leaving nurse will first acknowledge the patient that the incoming nurse will take over, and then talk to the incoming nurse about this patient's situation outside the room. This step largely constitutes the act of passing patient care information to the next shift nurse as a means to maintain continuous care for the patient. After the verbal handoff, the leaving nurse signs off on their patients to the incoming nurse via the EMR system. After the formal handoff, the four patients automatically appear under the in-coming nurse's name when they first sign into the EMR system. The formal EMR signoff transfers the patient records and the responsibility of patient care officially to the in-coming nurse, as well as pending orders and other unfinished tasks.

4.4 Three Types of In-shift Handoffs without EMR

Handoffs not only occur between two shifts; they also frequently occur during the same shift, as a way of maintaining continuous coverage in ED care. In-shift handoffs

¹ All the names used in this paper are pseudonyms.

are usually short-term, ranging from a few minutes to one hour. Unlike per-scheduled shift changes that occur in 12-hour intervals, in-shift handoffs are more spontaneous and rely on the interpersonal coordination among collocated nurses. This section describes how these in-shift handoffs occur and are coordinated on the ED floor.

Lunch breaks – Planned One-hour Coverage. Perhaps the most important task for a float nurse is to cover for lunch breaks, since nurses cannot work continuously for a 12 hours without taking a meal. “Lunch” is an hour-long time off in the middle of a shift, no matter if it is at midnight or in the early morning. Lunchtimes are pre-scheduled by the nursing manager and announced in the shift meetings. To maintain continuous coverage with only one float nurse, room nurses working the same shift are all assigned with different lunch times. As a result, lunch can happen anytime during a shift.

The float nurse moves to different working locales, covering lunch-breaks based on the nursing schedule. Although the lunch break is only an hour long, necessary patient information must still be handed over to a float nurse so that patient care tasks can be performed. During the handoff, room nurses go through their patients one by one, and remind the float nurse of tasks to be done in the following hour, *e.g.* the next two hour vital sign check up and monitoring for chest pains. This lunch hour handoff is similar to the “verbal handoff” that is performed at the end of shifts. Through this brief face-to-face conversation, the float nurse receives critical information about a patient and is given the key tasks that might be happening in the next hour.

Unlike end-of-shift handoffs, handoffs for lunch are transitioned only through verbal communication, but are not signed in to the EMR system. The consequence of this action is that while the room nurse is on break and is no longer the one caring for the patients, in the EMR system, the room nurse is still “in charge.” Orders signed from ED doctors are still routed to the room nurse, but not the float nurse who is caring for the room nurse’s patients. Since the EMR system is viewable to every ED staff, the float nurse could always check patient’s information in a patient’s records. Nevertheless, orders are not publicly accessible to everyone. However, what is justly notable in the lunch hour handoff is how the information and the responsibility of patient care are separated due to the lack of EMR-based signoffs. The lack of EMR based care transitions may result in serious consequences to medical work.

Ad Hoc Breaks – 10 Minutes Unplanned Coverage. The intensity of ED work may exhaust room nurses, potentially inducing errors and mistakes. Each room nurse is allowed to have approximately 3 ad hoc breaks per shift when and if they feel the need, each lasting for 10 minutes. While room nurses are not expected to take too many breaks, they are not expected to work during extreme tiredness either. The 10-minute absence from their patients’ bedside, though brief, still disrupts the continuous monitoring work that room nurses engage in, forcing them to find coverage for their ad hoc breaks via a float nurse. Similar to the lunch breaks described in the previous section, ad hoc breaks are negotiated verbally, and are not officially transferred in the EMR system.

The spontaneous nature of ad hoc breaks makes planning for them impossible, since the vagaries of tiredness are situated in the nature of the workday. This makes the coordination process of ad hoc breaks more challenging to perform in the ED. In an ideal situation, when a nurse wants to take a break, they should be able to contact

the float nurse right away and immediately arrange for them to cover their patients. In reality, ad hoc breaks are usually initiated by the float nurse due to the lack of communication among nurses.

Ad hoc breaks are managed through a *flow sheet* placed on the left nursing station in the main ED area, where everyone can see it. The flow sheet is a table-look-schedule that is filled out during the shift. On the flow sheet, there is one line for each ED nurse, with the first column showing the shift time of the nurse, followed with their name, lunch break time, three ad hoc break times, and a note section.

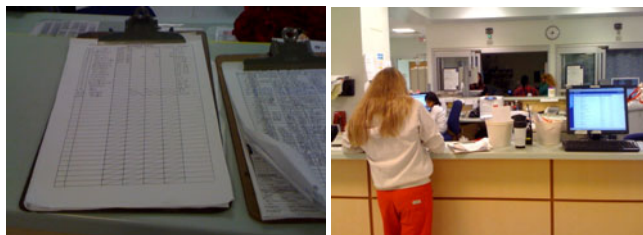


Fig. 2. The coordination of ad hoc breaks in the ED. Left, the flow sheet of nursing work; right, the float nurse is checking the flow sheet.

To coordinate ad hoc breaks, the float nurse checks the flow sheet constantly and speculates whether a room nurse may need an ad hoc break based on what has been documented on the flow sheet. For example, if nurse-A hasn't taken a break for the 4 hours since work began, and their lunchtime is schedule 3 hours later, the float nurse may consider giving nurse-A a break. Thus, the float nurse travels to nurse A's working locale and inquires whether a break is needed, and nurse-A will probably say yes given the circumstances. After nurse-A's coverage, the float nurse will sign the time of this ad hoc break on the flow sheet for nurse-A. In this way, the flow sheet plays an essential role as a central coordination artifact that allows the float nurse to manage continuous coverage. The ad hoc breaks are mainly initiated by the float nurse and the need for an ad hoc break is decided based on the flow sheet.

Use of the flow sheet to coordinate ad hoc breaks allows the float nurse to check, arrange and record for whom and when they have covered. The flow sheet greatly benefits the float nurse's work, but not those in need of breaks on the ED floor who are accidentally overlooked. On more than one occasion, we observed that nurses working in their locales were unable to locate the float nurse at the time they needed a break. Consider the following scenario²:

Mike, a room nurse stationed at one left corner of the main ED wanted to take a break. He is not feeling well today and has been working for 3 hours. He looked around every 5 minutes but didn't see the float nurse in the main ED area. Mike asked Lisa – the other nurse working in the same nursing station – whether she knows where the float nurse is. Lisa said, "Why don't you check the flow sheet and see where she is?" Mike ran to check the flow sheet on the other side of the main ED and was told by the nursing manager that the float nurse was covering a lunch break in

² Scenarios are all drawn from observation notes.

the urgent care unit. Mike came back and asked Lisa if she could cover him for a few minutes so he could get a snack at the fitting room. Lisa answered, "let me finish this vital first and I can keep an eye on [your patients] for you."

This scenario indicates the inefficiency afforded by coordination of ad hoc breaks using only the centralized flow sheet. Many times during observations, we saw the float nurse standing at the nursing station examining the flow sheet and considering who should be given a short break. The need for ad hoc breaks is not merely based on the dimension of time. Issues such as an individual's physical condition, the busyness of ED work, the unit a nurse is in and the demand of the nurse's patients also affects ad hoc breaks. Certainly, working at a trauma unit would be more exhausting than working at an urgent care unit, and night shifts produces more tiredness than day shifts. Unfortunately, time is generally the sole criterion for deciding whom to give ad hoc breaks in the ED.

When nurses can't find coverage from the float nurse, they have to ask for help locally due to their inability to travel far from their work locale. From our observations, nurses were able to ask for help from other nurses they could physically see, but were less likely to seek help from the other ED units, or even the other side of the main ED unit. Every nurse could use the phone at their ED station to call others stations, and even page the entire ED. The triage nurse has a walkie-talkie to communicate with the nursing manager. Yet, the float nurse has no communication devices on-hand and consensus seems to be that the overhead pages, which are intended for services such as paging doctors, should not be used for solving this seemingly trivial issue, such as breaks due to the tiredness.

Team Nursing – Shared High Workload Coverage. Though continuous coverage is viewed as linear temporal coordination in medical work [1], during this current study, we witnessed that continuous patient care are shared among multiple room nurses simultaneously, as developed by the nurse on the ED floor in the so-called, "Team Nursing" practice.

Traditionally, one ED nurse takes care of four patients. It is assumed that a nurse can manage the orders; check-up and monitoring work for four patients simultaneously. Sometimes though, this assumption becomes problematic, as there is no coverage for emergency situations. As one nurse told us, "*all of them [patients] are stable right now. No one knows what might happen [the] next minute; if two breakout at the same time, I am not gonna cover both of them at the same time.*" Here, even the room nurse is physically at the working locale, yet, attention can only be given to one patient at a time, leaving the other three patients momentarily uncovered. The general assignment of the ED would allow a nurse to cover four patients on a regular basis, but not when two or even more patients need to be attended to at the same time.

Helping with the room nurses' temporarily heavy workload is theoretically one of the daily duties of the float nurse. However, situations wherein the room nurse is physically present but is unable to give the needed attention to a patient tend to be urgent, and though the float nurse could be called, there probably would not be enough time for the necessary patient information to pass from the room nurse to the float nurse. In a new strategy developed by nurses at our field site called "team nursing," nurses cover locally for each other in cases such as these.

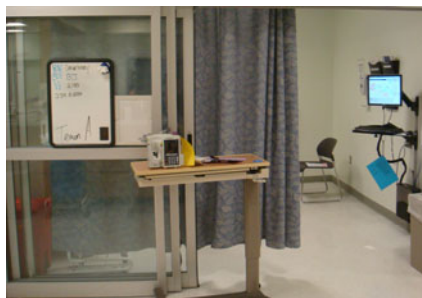


Fig. 3. The whiteboard hung outside each patient room. It shows the room nurse belongs to “Team A”.

“Team Nursing” pairs nurses into teams based on their physical proximity. This physical proximity allows nurses to cover each other when needed. In “Team Nursing,” however, nurses do not only provide coverage when urgent situations break out; instead, nurses are asked to stay aware of the situation of all the patients of all the nurses on their team. For example, the two nurses in Team A work in the left nurses’ station, and have eight patient rooms collaboratively managed between them. ‘Team Nursing’ is useful because, as one room nurse explained, *“if anything happens, we know why this patient is here and what happened to him, sometimes a few seconds can save a life.”* When all team members are constantly check of each other’s patients, the verbal handoff can be avoided when an urgent situation arises. In addition, team members are able to gain an understanding of all the patients in their team gradually and thoroughly over time – an understanding that is not possible for a float nurse to obtain even through a verbal handoff.

“Team Nursing” provides room nurses with a way to avoid possible information breakdowns during emergency situations. Nurses give and receive constant patient updates, and read the charts of each other’s patients. Through team nursing, nurses can avoid the common practice of verbally reporting the patient situation to the covering nurse during a time of urgency. As one of the room nurses we spoke with commented in the previous paragraph, a matter of one second can save, or lose, a life, and by shaving away the time a handoff takes, precious time is added to direct patient care. With more than one nurse paying attention to a patient, it is also likely that potential medical errors could be avoided through the overlapped attentions.

Nevertheless, in ‘Team Nursing,’ each nurse still receives their own assignments and there is no cross-coverage among these assignments. The ‘teams’ of nurses only exist on the whiteboard; nurses still act solely in the EMR system. What is shared in “team-nursing” is the responsibility and information of a pool of patients. Each nurse gains not only a sense of partial responsibility for all the patients of their team, but also for the nurses who are their team members. It is also notable that the creation of nursing teams is coordinated mainly through spatial proximity, and not a linear temporal arrangement.

5 Discussions

This section takes a synthesized view towards in-shift handoffs. First, we discuss how in-shift handoffs are managed through the interplay of temporal and spatial coordination on the ED floor. Second, we highlight the issues resulted from the lack of in-shift handoffs in system design.

5.1 The Coordination of In-shift Handoffs

Though the continuous coverage concept applies to nurses and doctors, in-shift handoffs only occur among nurses, and not ED doctors. In this section, we discuss how in-shift handoffs are coordinated by nurses collocated on the ED floor.

Working Locale Attachment. As evident in our study, in-shift handoffs apply solely to nurses, but not ED doctors. Doctors rarely seek coverage if they need to get lunch outside the ED or leave for 5 minutes. This is largely due to the nurses' attachment to a specific working area. The doctors' work locale, as described in the ED work section, is the entire ED. Doctors accept patients from any place in the ED and remain in the MD station for most of their shift. Because the doctor's work is not attached to a specific patient room, there is no such notion of leaving patients for a short break, and consequently, no need for in-shift coverage.

The strong attachment nurses exhibit to their working locales can even be seen in the way that their daily assignments are given; nurse are never assigned to patients, but room numbers. Spatially, the room nurses manage neighbor each other and the nursing station they use is right outside their patient rooms. This setup illustrates how nurses need to be physically present to maintain the continuous coverage of their patients - in other words, nurses have to be covered as soon as they step out of their designated working locales. In this sense, taking lunches downstairs at the cafeteria is an absence from work, as is drinking water at the locker room next door, or even going to restroom. All of these require nurses to find short-term replacements.

Only one of the three types of in-shift handoffs noted in this paper can be anticipated: lunch break handoffs. The other two forms of coverage, related to ad hoc breaks and high workloads, are more spontaneous in nature and are usually coordinated at the local level. Interestingly, no matter how long the coverage is, it inevitably involves handing patients from one room nurse to another, and the consequences for one 5-minute break is similar to a full shift handoffs. The current system design lacks sufficient acknowledgment of these short-term handoffs, potentially leading to other types of handoff breakdowns.

From Temporal Transition to Local Awareness. More importantly, the in-shift handoffs identified in this study are often coordinated locally by nurses working at nearby locales. This coordination of in-shift handoffs shows that the scheduling process of handoffs is not merely a temporally oriented task, but also one that takes place at the locations where nurses engage in work.

Intershift level coverage is coordinated through rotations of temporal rhythms in medical work [8]. Rhythms such as shifts follow a strict 12-hour timeline and are a linear process of one after another. There is generally no overlap between two shifts other than the brief transition period. In contrast, among the three identified in-shift

handoffs, only the lunch break can be scheduled since it occurs only once during the shift, and has to be spread out to different times (once for each nurse) for the sake of scheduling. The other in-shift coverage are more spontaneous in nature since it is not easy to predict when a nurse may need an ad hoc break due to tiredness or when an extra hand will be needed due to medical urgency. The more unpredictable a break is, the more unlikely it is to schedule the coverage in advance via temporal coordination.

As described in the findings section, ad hoc breaks are covered mostly through spatial arrangements, as opposed to temporal coordination. For ad hoc breaks, the nursing manager allows each nurse to have approximately 3 short breaks during their shift. Yet, the specific times that a nurse requires a break cannot not be scheduled, and the float nurse may not be available or be able to notified when a room nurse need it. As a result, nurses working close by may be asked to cover for a break when the float nurse cannot. This is an example of a local coordination mechanism between nearby working locales that enables continuous coverage in the ED.

In addition, coverage of urgent situations by “Team Nursing,” designed at the field site we worked at, also reaches beyond the linear temporal coordination that most studies previously reported. Nurses in a team cover patients not in the “one-leave-one-in” mode. By overlapping their coverage for a single patient, each nurse still has their own assignments, but stays aware of all the team’s patients so that in times of urgency, there is no need to go outside the team for coverage. This pattern switches the linear coverage into the paralleled overlap coverage. It is important to note that this parallel coverage is also enabled through the spatial coordination within the collocated nursing teams.

Just as Randell argues for the use of cognitive artifacts to help physicians stay aware of patients at all times, as opposed to handing jobs over just during the short handoffs periods, local awareness also facilitates the continuity of patient coverage and can alleviate the sudden “drop the baton” situations that can happen during transition moments [16]. In “Team Nursing,” nurses in the same team know each other’s patients naturally through their physical proximity. Many times in the study, we saw the nurses asking about the status of another team member’s patients, or even sharing their patient care information through casual chats at nursing stations. When this does not happen naturally, “Team Nursing” practice requires nurses in the same team to update each other on patient cases throughout the entire shift, and to constantly check up on each other’s patient charts in the EMR. So when it comes to managing in-shift handoffs through local coordination, it is likely that the nurses working nearby would already have certain prior knowledge about each other’s patients. Thus, the burden of in-shift handoffs is no longer in the short transition period, but accumulated over time through constantly updating and acknowledging patients information. From this sense, the local awareness is the knowledge of things going on in their local area of neighboring patients. Indeed, this local awareness is a continuous endeavor rather than a sudden transition. Even though the actual handoff may still happen in a short period of time, the information needed for patient care is updated gradually as a shift progresses.

5.2 The Impact of In-shift Handoffs

Much of medical work is now mediated through the use of the EMR systems. Consequently, the system also impacts how handoffs are conducted on the nursing floor. In this section, we discuss the potential impact caused by the lack of the proper EMR-based in-shift handoffs on the ED floor.

The Lack of System Supported In-shift Handoffs. Previously we used the definition of “transition of information, responsibility and authority” between two people to refer to handoffs in the medical field. Zerubavel [1] also emphasized the importance of responsibility in the handoff process. In previous studies, the layers of handoff, such as responsibility, information and artifacts, are transitioned simultaneously as nurses pass over to the next shift [5][6]. However, handoffs in the ED take place in two-phases: first, passing over authority and information at the patient bedside, and second, officially signing-off responsibility in the EMR at the nursing station. Both phases are essential for the quality of patient care and are necessary steps to go through, since the EMR is not only a patient record system, but also a document of work activities that could be for liability purposes *e.g.* [18][19].

Nevertheless, most short-term handoffs only go through the initial stage of verbal transitions, but not in the information system. This is partially due to a lack of design to support short-term coverage in the information system. As is, the handoff process now is so complex, nurses would assume it acceptable not to signoff, and continue to find a nurse to cover for a few minutes. Similarly, “Team Nursing” only exists on the whiteboard hanging outside the patient rooms, and not in the EMR system. In this case, responsibility of a patient still falls completely on the room nurse, despite the rest of “Team A” checking the patient’s chart in the EMR and caring for the patient when emergencies happen.

The lack of formal records for short-term handoffs bears various consequences for medical work. Nowadays, the EMR is not only a database of patient records, but serves as a central system of supporting all types of clinical work [20]. Evidence shows that the EMR is increasingly used to trace clinical activities for work efficiency, medical errors and responsibilities [18][9]. The lack of EMR-based transitions during temporary handoffs may lead to mismatch between the actual work practice and the records of work practices stored in the system. For instance, the covering nurse’s tasks will not be documented in the formal medical records. When these tasks are checked afterwards, the actual nurse who performed the tasks and the nurses who is indicated in the system as the task handler will not be the same person. Also, it is arguable that many breaks are not documented properly in the flow sheet since they are handled locally without leaving any records with the nursing manager. This is not only legally conflicting, but also challenging to the ED management since the manager is now oblivious to how many breaks a nurse takes in a shift – key information that needs to be kept track of at management level.

Illusion of Communication. Since the deployment of the EMR system, most patient care information is mediated through the information system. Orders from doctors are routed

to the “charge” nurse automatically.³ The charge nurse receives these orders as pending tasks whenever they sign in to the system. Due to the busyness and multi-patient care of ED practice, doctors may not be able to follow up nurses in person with each order they prescribed, but just assume that these orders will be checked and performed as the earliest convenience on the nurse’s end. Nevertheless, the lack of design for in-shift handoffs could potentially lead to delaying in receiving information in the ED.

Since orders are no longer passed along on paper or through phone calls, room nurses who are monitoring patients information have to check the EMR system frequently to see if there are any pending orders. As what we observed in the study, nurses always log into system every 2-3 minute to see if they have any new pending orders. This frequent checking ensures timely delivery of medications to avoid patient suffering and delay in care.

However, the lack of concern for in-shift handoffs may lead to delays in receiving patients’ orders. Doctors who order a test or new medication may not know that a nurse is on a designated lunch break at the time they put in an urgent order. Similarly, the float nurse may not be aware of a pending order since the order is only routed to the room nurse who is on break. This delay in receiving information has been referred to as *illusion of communication* in previous literature [21]. The illusion happens when one assumes communication is happening after orders are put into the system, but the orders may not be checked in the system by the other. This generally is not an issue for ED nurses, as they frequently log in and check the pending orders every a few minutes. But since the orders are not marked as pending orders in the float nurse’ to do list, it may be overlooked during the in-shift handoffs. From this sense, the illusion of communication is partially due to the lack of support for in-shift handoffs in the information systems.

6 Implications for System Design

The findings of this study suggest that continuous coverage should be considered a primary design principle for clinical systems. Drawn from our qualitative field observations in the ED, we found that continuous coverage is maintained not only through the handoffs between two shifts, but also through various non-working moments that occur during the same shift of work. The lack of a proper design for these in-shift handoffs may impact both nursing work and the nursing/doctor collaborations that are mediated through the information systems. In this section, we outline possible design opportunities based on the findings of the study. Noteworthy is that what we suggest here is by no means the only design solution. Rather, it is an implication leading to the designs of clinical systems that take in-shift handoffs into considerations.

6.1 Easy Role Switch During In-Shift Handoffs

As described earlier, only at the end of shift do nurses sign-off patients officially in the EMR system; other in-shift handoffs are conducted verbally. Indeed, if the

³ This implication is based on the observations from the current study. This may not be applied to other EMR systems or other emergency departments.

in-shifts handoffs were handled in the system, nurses would have to sign patient to the covering nurse first, then sign it back a few minutes later. The complexity and extra procedures involved in the EMR-based signoff keeps nurses from doing it during temporary breaks. This suggests that end-of-shift handoffs and in-shifts handoffs should be handled differently in the EMR system. System design should allow for easier role transition during temporary coverage periods, such as providing a *covering* feature to allow a one-click temporary signoff during in-shift handoffs, or add float nurses as a *covering person* in the system and allow role transitions for the one being covered directly. Simplifying the system signoff procedures would encourage clinicians to use it for the brief in-shift handoffs process and to avoid the potential communication illusions and delays.

6.2 Social Display of In-shifts Handoffs

Clearly indicated in our study, ED clinicians work in different working locales without much overlap. Room nurses stay nearby the patient rooms they manage, and ED doctors work primarily at the MD station. Nevertheless, the status of nurses and the orders doctors prescribed have to be coordinated during the in-shift handoffs. We suggest the EMR system to incorporate a social display system in order to provide awareness information for clinicians working on the ED floor. The display system can mark the status of each nurse and new orders for each room using various color schemes on a shared ED floor map. This map is then shown on the EMR front page or even projected onto a larger screen that can be publicly seen on the ED floor. In doing this, everyone in the ED can be aware of each other's working status, like who is on break (*e.g.* in yellow color) and which patient has new orders (*e.g.* in red color). In this case, the illusions of communication can be eliminated since new orders will be visible without logging onto the EMR system. Doctors could also be cautious when they notice the room nurse is on break. Nurses can also remind each other as the colors associated with patient room is publicly displayed. Using the display system can benefit the communication between doctors and nurses in general, since by flagging new orders in the social display system, room nurses no longer need to log in to the EMR every 2-3 minutes to see if they have pending orders.

6.3 Shared Responsibility

Different from the temporal work coordination that occurs at the end-of-shift handoffs, in-shift handoffs are often managed through spatial coordination among nearby nurses. That is, nurses working nearby act as a team to help each other when necessary. The local coordination turns the solo practice of "one nurse cares for four patients" into a collaborative practice of "three nurses in team A manage twelve patients together." To do so, nurses in a team maintain a local awareness by constantly checking on and updating patients' situations with each other. We suggest the design of patient care systems take advantage of local awareness and provide shared responsibilities for nurses working nearby. This way allows team nursing to be formalized in the medical records. Team members could not only have access to patient information, but are also able to document and order in the records. In this case, the medical record become a cognitive artifact [16] that provides constant

information sharing for in-shift handoffs and relieves the possible “baton” drop transition moments.

7 Conclusion

In this study, we explore how the concept of continuous coverage is practiced in an emergency department. Our observations reveal patterns of unreported short-term handoffs in the same shift of nursing care. These in-shift handoffs result from the strong attachment of nursing work to the working locale they are assigned. Differing from previous studies on handoffs, this current study shows that when handoffs occur within the same shift, they are more likely to be coordinated through local arrangement among collocated nurses. The handoffs that are arranged through the closeness of working locales enable constant updates among nurses and provide them a sense of local awareness – an awareness of knowing patients situations in the nearby rooms. This mechanism could avoid the sudden breakdown in the normal handoff processes. However, since these short term handoffs are hidden in the same shift work and are not recognized by system designers, they may lead to issues such as the separation of information and responsibility in patient care, and the illusion of communication, if they are not properly handled. The findings of this study call attention to the design of systems that account for in-shift handoffs that can improve handoff processes via the coordination of local awareness during ED work. In-shift handoffs may also exist in areas that require strong attachment to an employee's working locale, such as space shuttle, software design, and 24/7 services areas.

Acknowledgements. We would like to thank Drs. Harris R. Stutman, Giancarlo P. DiMassa and ED staff at Long Beach Memorial Medical Center. Thanks to Victor Ngo for helping proofread the earlier draft of this paper.

References

1. Zerubavel, E.: *Patterns of time in hospital life: a sociological perspective*. University of Chicago Press, Chicago (1979)
2. TeamSTEPPS.: *Team Strategies and Tools to Enhance Performance and Patient Safety: Pocket guide (AHRQ Pub No 06-0020-2)*. Agency for Healthcare Research and Quality, Rockville, MD (2006)
3. Sharma, N., Cohen, M., Hilligoss, B., Patterson, E.: *Handoffs & Handovers: Collaborating in Turns*. In: *CSCW 2010 Workshop*, Savannah, Atlanta (2010)
4. Wilson, S., Galliers, J., Alem, L.: *Handover: Collaboration for Continuity of Work*. In: *ECSCW 2007 Workshop*, Limerick, Ireland (2007)
5. Tang, C., Carpendale, S.: *An Observational Study on Information Flow during Nurses' Shift Work*. In: *Proceedings of CHI 2007*, pp. 219–228 (2007)
6. Zhou, X. Ackerman, M.S., Zheng, K.: *I just don't know why it's gone: maintaining informal information use inpatient care*. In: *Proc. CHI 2009*, pp. 2061–2067 (2009)
7. Berg, M.: *Accumulating and Coordinating: Occasions for Information Technologies in Medical Work*. *Computer Supported Cooperative Work* 8(4), 373–401 (1999)

8. Reddy, M., Dourish, P.: A finger on the pulse: Temporal rhythms and information seeking in medical care. In: *Proceeding of CSCW 2002*, pp. 344–353 (2002)
9. Bardram, J.E., Bossen, C.: *Mobility Work: The Spatial Dimension of Collaboration at a Hospital*. In: *Proceedings of the ACM CSCW 2005*, pp. 131–160 (2005)
10. Patterson, E.S., Roth, E.M., Woods, D.D., Chow, R., Gomes, J.O.: Handoff strategies in settings with high consequences for failure: lessons for health care operations. *Int. J. Qual. Health Care* 16, 125–132 (2004)
11. Friesen, A.M., White, V.S., Byers, F.J.: Handoffs: Implications for Nurses, in *Patient Safety and Quality: An Evidence-Based Handbook for nurses*. AHRQ (2008)
12. Gandhi, T.K.: Fumbled handoffs: One dropped ball after another. *Ann. Intern. Med.* 142, 352–358 (2005)
13. Ebright, P.R., Urden, L., Patterson, E.: Themes surrounding novice nurse near-miss and adverse-event situations. *J. Nurs. Adm.* 34, 531–538 (2004)
14. Haig, K.M., Sutton, S., Whittington, J.: SBAR: a shared mental model for improving communication between clinicians. *Jt Comm J. Qual. Patient. Saf.* 32, 167–175 (2006)
15. Sarcevic, A., Burd, R.S.: Information handover in time-critical work. In: *Proceedings of the GROUP 2009*, pp. 301–310 (2009)
16. Randell, R., Wilson, S., Woodward, P., Galliers, J.: Beyond handover: supporting awareness for continuous coverage. In: *Cognition, Technology & Work* (2010)
17. Strauss, A., Fagerhaugh, S., Suczek, B., Wiener, C.: *Social Organization of Medical Work*. University of Chicago, Chicago (1985)
18. Ash, J.S., Berg, M., Coiera, E.: Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J. Am. Med. Inform. Assoc.* 11, 104–112 (2004)
19. Koppel, R., et al.: Role of Computerized Physician Order Entry Systems in Facilitating Medication Errors. *JAMA* 293, 1197–1203 (2005)
20. Heath, C., Luff, P.: Documents and Professional Practice: ‘bad’ organizational reasons for ‘good’ clinical records. In: *Proc. CSCW 1996*, pp. 354–363 (1996)
21. Dykstra, R.: Computerized physician order entry and communication: reciprocal impacts. In: *Proc. AMIA Symp.*, pp. 230–234 (2002)

A Taxonomy of Microinteractions: Defining Microgestures Based on Ergonomic and Scenario-Dependent Requirements

Katrin Wolf¹, Anja Naumann¹, Michael Rohs², and Jörg Müller¹

¹ Deutsche Telekom Laboratories, TU Berlin, Ernst-Reuter-Platz 7,
10587 Berlin, Germany

² LMU Munich, Amalienstr. 17, 80333 Munich, Germany

katrin.wolf@acm.org, {anja.naumann,joerg.mueller03}@telekom.de,
michael.rohs@ifi.lmu.de

Abstract. This paper explores how microgestures can allow us to execute a secondary task, for example controlling mobile applications, without interrupting the manual primary task, for instance, driving a car. In order to design microgestures iteratively, we interviewed sports- and physiotherapists while asking them to use task related props, such as a steering wheel, a cash card, and a pen for simulating driving a car, an ATM scenario, and a drawing task. The primary objective here is to define microgestures that are easily performable without interrupting or interfering the primary task. Using expert interviews, we developed a taxonomy that classifies these gestures according to their task context. We also assessed the ergonomic and attentional attributes that influence the feasibility and task suitability of microinteractions, and evaluated their level of resources required. Accordingly, we defined 21 microgestures that allow performing microinteractions within a manual, dual task context. Our taxonomy poses a basis for designing microinteraction techniques.

Keywords: gestures, microinteractions, dual-task, multitask, interruption.

1 Introduction

Human-computer interactions are to a great extent defined by hardware design such as the size limitations and the interconnections of the hardware components. For instance, the size of current smart phones is mainly determined by the screen size necessary for watching multimedia content or browsing the internet.

Novel concepts of interaction design and HCI research tend to split the interface into specialized components, especially for separating the hardware that processes the user input [5, 8, and 15]. For example, Loclair [8] uses a depth camera for tracking pinch gestures; Harrison [5] measures body transmitted acoustic signals that are generated by tapping a finger against other fingers or the forearm; and Saponas [15] is using EMG to recognize finger pressure and finger taps. These works focus on the input and sensing

techniques for tracking hand gestures for microinteractions. Microinteractions, defined by Ashbrook as short-time interruptions of primary tasks [1], can have huge benefits in allowing mobile application control in parallel to ongoing primary tasks and could significantly expand the set of tasks we could perform on-the-go. Chewar [2] defined secondary tasks as those which can take place concurrently with the primary task. However, there is a research gap in investigating microinteractions from the task-driven perspective and from the human point of view [17].




We understand microinteractions as interactions that are task-driven and goal oriented, and which may include system feedback. They can be evaluated with traditional usability metrics such as effectiveness, efficiency and user satisfaction. In contrast, microgestures are actual physical movements, e.g. of fingers, which are recognised by the system, and where the system reacts upon. Microgestures are part of microinteractions. Within the related work of microinteractions, the main focus is on short-time manual motor interruptions, or on manual synchronous tasks. We investigate microinteractions that can be performed synchronously. The attentional resources then have to be used alternately or in parallel.

This paper explores and identifies microgestures and finger movements that are performable and does not draw significant attention away from the primary manual task which is to be done in parallel. In deciding the manual primary task, we focused on manual grasp research that is done in the rehabilitation and medical science areas.

Feix [4] developed a grasp taxonomy that compared 14 grasp taxonomies based on 92 years of human hand's research. He identified 33 different human natural grasps and classified them into 3 main types: palm, pad, and side. We abstracted this taxonomy and related it to our research interest: microgestures performed alongside manual tasks (see Table 1). The left three columns of the table shows the original main grasp types of Feix' taxonomy and describes one specific example for each type. The right column shows which free movement potentials we identified for the taxonomy's main grasp types. For investigating microinteractions that are meant to be executable alongside manual tasks, we have chosen 3 exemplary tasks: each one is using one grasp of one main group of Feix' taxonomy. Thus, we aim for ensuring research results that are scalable to a wide range of manual activities.

Primary tasks, such as driving a car or holding objects, do not need our complete cognitive attention nor are all fingers strictly involved in these processes. This allows for performing a second task at the same time. This task can be related to a different context like answering the phone while driving a car. Alternatively, controlling mobile applications by microinteractions could also offer the opportunity to apply subtasks through adding augmented function to the primary task without interruption. For instance, the input for many mobile applications in the automotive context, such as setting up the navigation system, controlling the music player, or opening and shutting the car windows, could be realized by microinteractions that are performable without releasing the steering wheel and therefore not interrupt the manual effort of the primary task.

Table 1. Microgesture options during ongoing manual tasks: Analysis of Feix’s grasp types: Palm, Pad, and Side, into which all human grasps can be categorized. Fingers are counted starting from the thumb.

Grasp type (Feix [4])	Description (Feix [4])	Involved hand-parts (Feix [4])	Potentially still movable hand-parts
PALM (e.g. Steering a car)	 Medium wrap	Low power grasp performed by 2-directional force between palm (finger 2-5) and abducted thumb	Particular fingers and thumb
PAD (e.g. Inserting a cash card into an ATM)	 Precision grasp	2-directional force between abducted thumb and index finger	Finger 3-5: middle, ring, and little finger
SIDE (e.g. Drawing with a stylus on a graphic tablet)	 Dynamic tripod	2-directional force between: a) added thumb and middle finger while index finger stabilizes or b) thumb and index finger while middle finger stabilization	Ring, little finger Stabilizer: index finger or middle finger

For the palm grasp for example, we have chosen driving a car as primary task that allows microgesture commands such as tipping or dragging at the steering wheel (see Fig.1).



Fig. 1. Performable microgestures while steering a car with a palm grasp: Tipping fingers on the wheel or dragging it with the thumb

In contrast to the chance of enriching the primary task conceptually by allowing a secondary task to be performed simultaneously; there is a risk that the performance of the primary as well as of the secondary task might decrease, because of attentional deficit. [3, 20].

2 Related Work

We relate our work to research that is investigating microinteractions performed by hand gestures. We focus on the effect of multitasking on motor and attentional efforts, on gesture-based interaction techniques as well as on wearable gesture tracking systems that do not limit the hand skills like data gloves do by reducing the touch sensitivity of the hand.

Within the human-factors related research, multitasking is investigated focusing on task interruptions and attentional issues of both: the primary and the secondary task. While Wexelblat [19] and Quek [13] claim that gestures are not natural for computer interactions because they only represent a small part of human communication, Karam [7] suggests that this “small part” is potentially well matched to secondary interactions. McCrickard [9] investigated the effects of distraction and recovery caused to a primary task (editing a text document) by a secondary task interruption, which was a notification for receiving an instant message. For the specific case of dual-task-microinteractions, there is a gap of research about how to design dual-task scenarios and how to select microgestures. For keeping the performance stable, there are two strategies: alternating two tasks or performing them in parallel. This is possible if at least one task can be performed with a certain level of automation and therefore requiring limited attention.

Wickens’ Multiple Resource Theory (MRT) Model describes that two actions can be done in parallel if at least one has reached an automated level through learning [20]. Based on Wickens’ Multiple Resources Theory [20], Oulasvirta [12] developed the Recourse Competition Framework (RCF). He is investigating cognitive resources when users are on the move. Oulasvirta explains that the resources are partly reserved for passively monitoring and reacting to contexts and events, and partly for actively constructing them. This model suggests that the resources for competitive task interactions alternate through breaking down the primary fluent interaction for up to four seconds.

Another research field that concerns about microinteractions and their trackability and classification is computer science. Computer vision based gesture tracking for identifying pinch gestures has been investigated by Loclair [8]. Vardy [18] tracks finger flexion with a camera integrated in a wrist band. Howard [6] uses optical detectors (that are also integrated in a wrist band) for measuring LED light that is reflected by the fingers. Harrison [5], Saponas [15], and Rekimoto [14] measure hand gestures using body transmitted signals, such as acoustic signals, EMG, and electrodes that display forearm movements by capacitive sensing.

So far, several multitasking scenarios and interaction techniques have been explored and tracking technologies for microinteractions have been developed and evaluated. But there is still a research gap in classifying microinteractions regarding their ergonomic dual-task potential. We investigate which microgestures might be best suited when applying secondary tasks in addition to certain exemplary primary tasks. Therefore we aim to develop a taxonomy based on fundamental ergonomic and anatomic hand research [4]. Our taxonomy can serve as a basis for developing novel microinteraction-based interfaces.

3 Method

The goal of our study was to generate taxonomy for microinteractions by listing and evaluating all microgestures that are performable alongside the main grasp types. The taxonomy aims to develop a general hand gesture set as well as to display ergonomic issues related to hand gesture performance, the necessary attention to perform the gestures, and the risk that the gesture is performed unintentionally as a natural movement and therefore would be misinterpreted as an input command.

A common method for defining gestures in the HCI field is to involve users in the design process [21]. To create a gesture set that already contains gestures of good feasibility and to generate valid data about how the majority of the users will be able to perform these gestures while continuing a manual task, we decided to involve experts, who know about the motor abilities and limitations of the majority of the users. Therefore we interviewed one sports therapist and three physiotherapists separately and asked them to evaluate a gesture set using props (see Fig. 2) regarding ergonomic and scenario-related aspects as well as to find more gestures that might suit the use case.

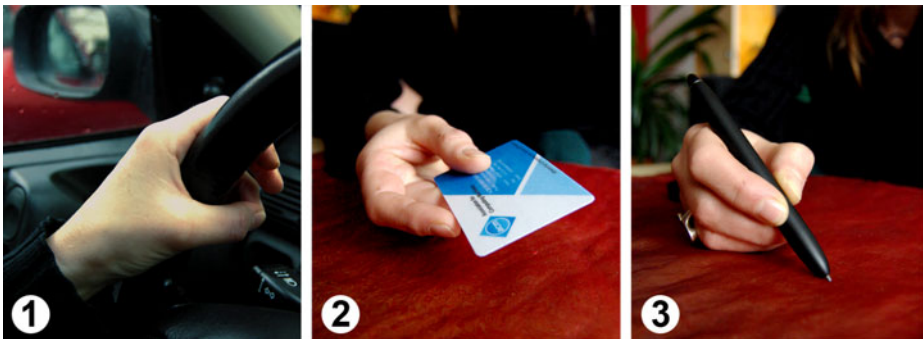


Fig. 2. The participants are testing the feasibility of hand gestures while (1) holding a steering wheel (2) targeting a cash card, and (3) drawing with a pen

We interviewed the experts separately in two sessions (see Fig. 3). We started the first session with a prepared set of 11 hand gestures, which were graphically presented to be evaluated by the experts. This initial gesture set consisted of seven palm-gestures, two pad-gestures, and two side-gestures, which were already used within microinteraction research projects [1, 5, 6, 8, 14, 15, 18]. For each gesture, we asked the experts to evaluate its performance ability by answering the following questions:

Feasibility. How easy is the hand gesture performable regarding ergonomic aspects when it is done eyes-free?

Limitations. Which ergonomic aspects limit the hand gesture performance?

Attention. Does the pure gesture performance require low, medium, or high attention?

Risk of confusion with natural movements. Could the gesture be randomly performed during the task?

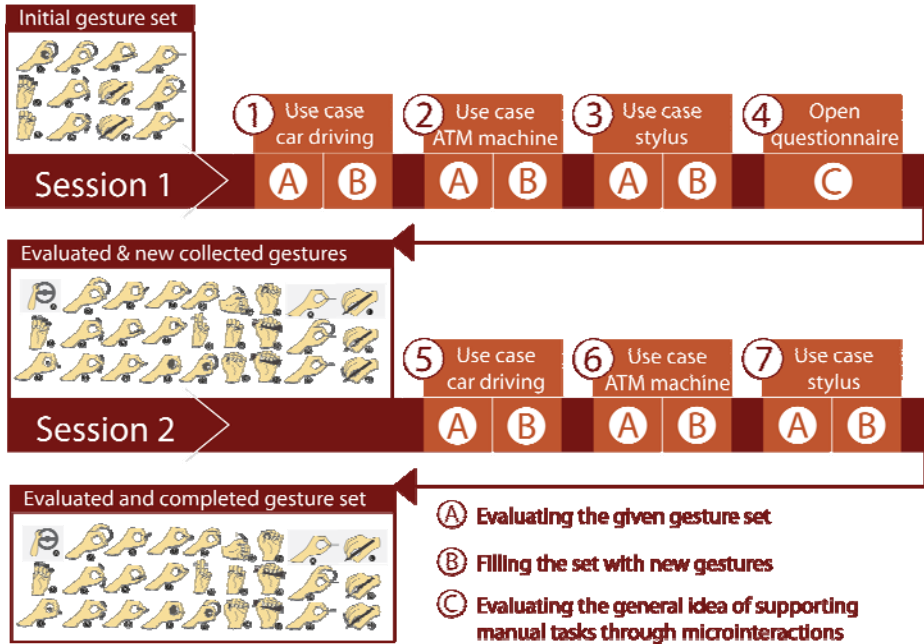


Fig. 3. Experiment walk through: We interviewed the experts separately in two sessions. The first started with a pre-defined gesture set, which the experts evaluated and completed using props. An open questionnaire completed the first session. Within the second session, the new collected gestures were evaluated by the experts by walking through the use cases with help of props again. The result was a completed and evaluated gesture set that shows feasible hand gestures that can be performed while continuing a grasp-based task.

For evaluating the different performance parameters in phase 1-3 and 5-7, we used different scale ranges: For the feasibility, we asked them to distinguish between easy (+) and hard (-). The required attention was valued at “low”, if the gesture execution was easily performable without influencing the main task performance. The value was “medium”, if the gesture execution required some of the attention away from the main task. The value was “high” if executing the gesture needed visual attention or if the main task might be interrupted. Within the evaluation section, we took notes of the verbal comments. Within the creation section, we took photos and drew sketches of the gestures the experts were performing.

After all evaluations of given gestures in one session, we asked the experts within the creation section to describe and perform further gestures that suit the specific context. We took pictures of these new identified gestures and added them as a graphical presentation to the gesture set for the next interview session.

The first sessions finished with an open interview about the experts’ general opinion about the idea to support a manual main task through microinteractions.

4 Results

The outcome of our iterative interviews was a list of 21 expert evaluated microgestures: 17 palm-, 2 pad-, and 2 side-gestures, as shown in figure 4 and described in greater detail in table 2.

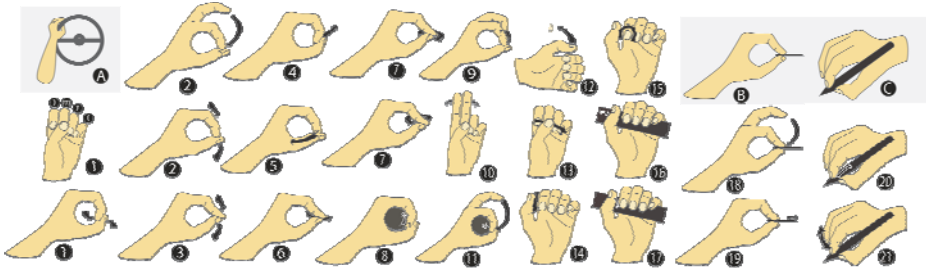


Fig. 4. The expert-defined and evaluated hand gesture set. The experts found 17 gesture types for the driving scenario (A). The card targeting scenario (B) and the stylus scenario (C) just contain 2 gesture types each. Most gesture types have several sub-types by performing them with different fingers (index, middle, ring, and little finger). Moreover the same gesture results in a different sub-type (e.g. touch, tab, or press), if it is performed with different acceleration or duration (see Table 2).

The very similar evaluation data of the different interviews regarding the valuation of the microgesture's required motor and attentional effort, allowed for comparing and concluding the results into one single table (see Table 2). The opinions we collected during the interviews are subjective expert arguments. In case there were different opinions about the feasibility or attentional efforts of a microgesture, we chose the more negative ones in order to exclude the less feasible gestures from further examination, and to make sure that the taxonomy will work for a large number of users. In the following, the results in Table 2 will be described in detail.

Arguments for valuing ergonomic issues were classified in sub-clusters: feasibility arguments that described why some gestures were hard or impossible to perform (limitations because of the shape of the grasped object or because of the anatomy of human hand). We identified arguments which described how well the primary and the secondary task fit together into a situation with simultaneously performed tasks. Within this category, we asked in particular for two aspects: attention and risk of confusion. The attention concerning comments describe if the in parallel performance of certain gestures requires high or low attention. The risk of confusion comments value the risk that a gesture is performed randomly as natural gesture or movement.

4.1 Feasibility and Limitations

We asked the experts to show us feasible hand gestures. In some cases, certain gestures have circumstance-dependant feasibility. For instance, the feasibility of touching, pressing, and tapping the fingers on the thumb while holding a steering wheel is dependant on the finger length and the wheel diameter (see Fig. 5).



Fig. 5. shows the feasibility of the third gesture of table 1: The thumb can be tapped easily with the middle (2) and the ring finger (3) while holding a steering wheel. But depending on the wheel diameter tapping the thumb with the index (1) or the little finger (4) can be difficult, especially for people with small hands.

There are mainly two classes of limitations in regard to the feasibility of microgestures. On one hand, the limitation is related to the physical objects that are to be grasped, for example, the size of the diameter of a steering wheel. On the other hand, feasibility is also limited by biomechanics, for example, it is difficult to move one finger without slightly moving its neighboring fingers as well.

There was a significant difference in feasibility between the index, middle, ring, and little finger, while performing some hand gestures, such as tapping a single finger on the thumb (Tab. 1, gesture 3). All experts were sure that the majority of the users will be able to perform an index-finger tab without any problems. Also, to move the little finger separately from the others was not a problem at all. The flexibility of the middle finger was a bit worse than of the index finger, but it was still feasible. However, the ring finger is always difficult to stretch separately. The degree of inflexibility varies individually; but the ring finger is considered to be the least feasible.

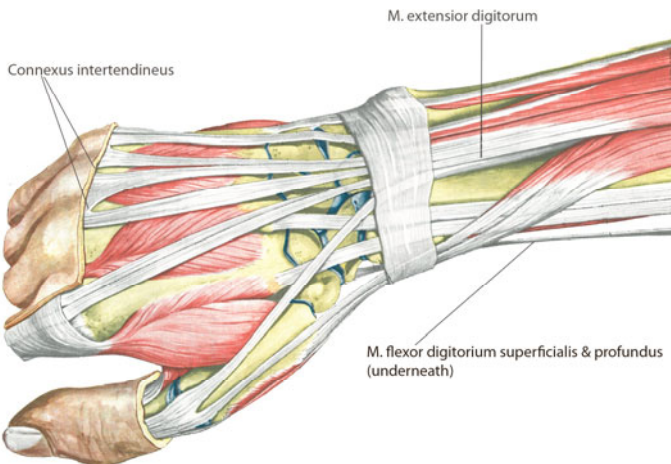


Fig. 6. shows the anatomic connection between the fingers that is responsible for the separation problem of the ring finger. Fig. 6 is a simplification of a figure in Spalteholz' Anatomy of Human [16].

The sports scientist expert explained this motor limitation and defined anatomic reasons like the connection between our muscles, sinews, and the fingers (see Fig. 6). Humans have more than 40 muscles to move the arm, hand, and fingers. If we aim to stretch the ring finger out from a palm grasp; two muscles (*M. flexor digitorum profundus* & *M. flexor digitorum superficialis*) are bending synergistically the index, middle, and little finger to bring them into the palm position. In addition another muscle is responsible for stretching the ring finger (*M. extensor digitorum*) but because this muscle is also responsible for stretching the other fingers and because the ring finger has a physical connection to the middle finger (*Connexus intertendineus*), the middle finger will always move a bit in the same direction as the ring finger does. The little and the index finger are more independently movable because they have their own muscles for stretching.

This means that in designing microgestures, it is preferable to focus on the index finger. In the case that a microgesture involves the ring finger, we will need to design it bearing in mind that the little and the middle finger will move slightly as well (see Fig. 7).

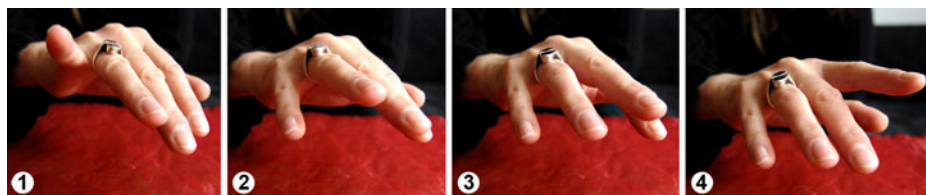


Fig. 7. shows the difficulty of stretching the ring (2) and middle (3) finger separately. Stretching the little (1) or the index (4) finger is much easier because of human hand's anatomic architecture which is shown in greater detail in Fig. 6.

4.2 Risk of Confusion

If commands are released by body movements, there is a risk that subconsciously executed natural movements can be misinterpreted as commands. For example, tabbing the steering wheel while driving a car (tab. 1, gesture 16) is a common behavior while waiting at the crossroads or listening to music. Reaming the thumb against the index finger would be expected while cooking, eating or putting salt on food, but while driving the risk of a ream-gesture occurring as a randomly executed natural movement is expected to be low.

4.3 General Idea

Besides gesture evaluation, we also did a questionnaire and collected the verbal comments of the experts on the general idea of allowing a secondary task alongside a continuous primary one. The opinion about the benefit of performing two tasks in parallel was different from one scenario to another. All experts think there is a huge benefit in being able to control a secondary task while driving a car. An example that is often used to support this argument is that drivers are anyway performing secondary tasks such as setting up the navigation system, controlling automotive

functions, or using mobile devices like cell phones while steering a car. The concept of controlling these devices or applications without releasing the steering wheel was valued positively for security arguments. The scenario of performing hand gestures while inserting a cash card into an ATM was not liked at all. None of the experts thought in parallel tasks could have a benefit for this use case. The last scenario about pen computing (e.g. drawing with a pen- or stylus-like input device on a graphic tablet) was modified during the interviews. Three of the experts thought that the possibility to change the stroke width or the color while drawing would have a bad effect on the precisions of the primary task but all of them said that having these options during short time interruptions could benefit the primary task. The flow of drawing would not be interrupted and therefore the task could be designed to be more comfortable than if a color selection would have to be done by keyboard or button-selection.

In general, the experts think palm grasp tasks best suit dual-task scenarios because these tasks are often low precision tasks and therefore require lower attention than pad or side grasp tasks.

Table 2. Microinteraction taxonomy. I =Index Finger, M=Middle Finger, R=Ring Finger, L=Little Finger, Th=Thumb, +=easy, -=difficult.

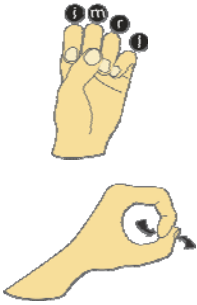

Gesture	Action	Ergonomic	Scenario compatibility
Palm-grasp gestures			
(1) 	(a) Tab	Feasibility Index (I):easy+ Middle (M): + Ring (R): + Little (L): diff. - Limitation By relation of finger length and hold object diameter, i.e. steering wheel	Attention Low: Th (thumb), I, M, L High: R, M Risk of confusion Risk to be a randomly performed natural move: high
	(b) Touch		
	(c) Press		Attention Higher than Touch-gesture, pressure rate is hard to control Risk of confusion High
(2) 	(a) Tab	Feasibility I: +, M: +, R: +, L: - Separation - : M+R Limitation By holding object diameter	Attention Higher than (1); Hard to distinguish from (3) Risk of confusion High
	(b) Touch		
	(c) Press		

Table 2. (Continued)

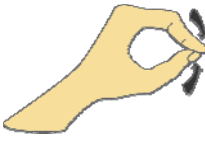
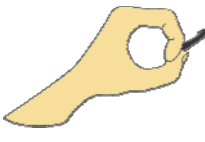
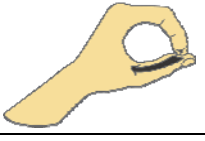
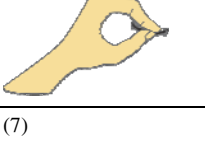
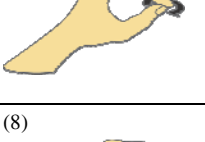
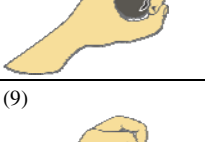

Gesture	Action	Ergonomic	Scenario compatibility
Palm-grasp gestures			
(3) 	(a) Tab	Feasibility M, R: + I, L: -	Attention Higher than (1); Hard to distinguish from (2) Risk of confusion High
	(b) Touch	Separation -: M+R	
	(c) Pinch	Limitation Object diameter	
(4) 	Flip	Feasibility I: +, M: +, R: +, L: - Separation No problem Limitation Object diameter	Attention Low Risk of confusion Low
(5) 	Drag&Drop index on thumb	Feasibility Just partly possible because of object diameter	Attention Medium Risk of confusion Medium
(6) 	Ream	Feasibility I: +, M: +, R: +, L: - Limitation hold object diameter (L)	Attention Low Risk of confusion Low
(7) 	Circle sidewise	Feasibility I: +, M: +, R: +, L: - Separation No problem Limitation Object diameter	Attention Individually different (+, -) Risk of confusion Low
(8) 	Drag fingers around the wheel	Feasibility -	Attention Medium Risk of confusion Medium
(9) 	Drag&Drop middle on index	Feasibility To complicated	Attention High

Table 2. (Continued)










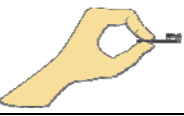

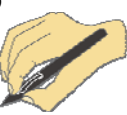
Gesture	Action	Ergonomic	Scenario compatibility
Palm-grasp gestures			
(10) 	Snip	Feasibility +	Attention Low Risk of confusion Low
(11) 	Tap the wheel	Feasibility I-L: +, Th: -	Attention Low Risk of confusion High
(12) 	Thumb up	Feasibility +	Attention Low Risk of confusion Low
(13) 	Drag&Drop thumb on finger nails	Feasibility Over I, M, R: + L: - Limitation Object diameter	Attention Low Risk of confusion Low
(14) 	Drag&Drop thumb on index-side	Feasibility Just partly possible because of object diameter Limitation Object diameter	Attention High Risk of confusion Medium
(15) 	Circle clockwise & contra-clockwise (CW & CCW)	Feasibility I, M: +, R, I: - Limitation Object diameter	Attention Individually different (+, -), but high for CW- / CCW-distinguishing Risk of confusion Low
(16) 	Drag thumb along object	Feasibility +	Attention Low Risk of confusion Low

Table 2. (Continued)

Gesture	Action	Ergonomic	Scenario compatibility
Palm-grasp gestures			
(17) 	Drag thumb around object	Feasibility -	Attention Medium Risk of confusion Low
Pad-grasp gestures			
(18) 	Tab	Feasibility I, R:- M, I, M&I: +	Attention High
(19) 	Drag middle finger above object	Feasibility +	Attention High
Side-grasp gestures			
(20) 	Tab I or M on object	Feasibility I. While drawing: - II. While holding: +	Attention I. High II. Low Risk of confusion I. High II. Low
(21) 	Drag Index or Middle finger on stylus up / down	Feasibility While drawing: - While holding: +	Attention I: Low, M: High Risk of confusion Low

5 Discussion

The microinteraction taxonomy shows that the design of microgestures, as well as their evaluation concerning usability related issues (e.g. ergonomic issues and scenario compatibility), is extremely dependent on the use context. This defines the primary task and rules the choice of the grasp type that is used to solve this task. The static gesture design as well as its feasibility (see table 1, column 1 & 3), is mainly influenced by grasp-related options such as hand anatomic limitations, but also ergonomic issues that are defined by objects and the character of grasping of the primary task. For instance, as explained below in greater detail, a low power palm grasp that is mostly used while driving, allows a lot of microgestures because releasing a finger from the steering wheel does not interrupt the task. Moreover, the primary task determines the attentional resources that are available to perform secondary task commands realized by microgestures.

5.1 Palm-Grasp Gestures

A low power palm grasp gesture allows for a great number of simultaneously performed microgestures without releasing the grasp. Palm related primary tasks that have a long duration require little attention by becoming an automatically performed process and leave a large part of the hand resources quite uninvolved. Thus, low power grasps seem well suited to be augmented by a large variety of microinteractions. Depending on the character of the primary task, some microinteractions have a high risk of being performed unintentionally during the primary task. Tapping on the steering wheel could be done while listening to music and drumming fingers on the wheel. To differentiate natural movements from input commands, three opportunities are possible for generating a gesture set:

1. Using a push-to-gesture event for telling the system that the parallel or subsequent movement is an intentionally performed command.
2. Designing commands as a combination of two gestures for reducing the chance of performing this couple unintentionally.
3. Defining design styles, e.g. rhythmic pattern, based on movements which are usually not done naturally in the primary-task-related context.

5.2 Pad-Grasp Gestures

Pad-grasp primary tasks such as inserting a cash card into an ATM machine due short//card slot?, use the 2 directional finger-thumb-force permanently, and require a high level of precision and short-term concentration. This was shown by our expert through demonstrating the failed attempt to perform both tasks in parallel. An added microinteraction would require interrupting or slowing down the primary task for a short time while performing the microgesture. According to the expert opinion, the interruption of the precisely short-term primary task is not acceptable because performing hand movements quickly and accurately does not allow microinteractions in parallel. Any finger movements would disturb targeting the cash card into an ATM by dismissing the target or extending the targeting time. Targeting and performing microgestures at the same time without risking high error rates on one or even both tasks is not possible. Moreover, the available hand resources for performing microgestures while interrupting the pad-grasp but still holding the tool are very limited.

5.3 Side-Grasp Gestures

Performing microgestures alongside a side-grasp drawing is hardly possible. Drawing is a highly precise manual task which is built on accurate hand movements and does not allow for the moving of fingers at the same time without having a negative effect on the quality of drawing. However, brief interruptions (to stop drawing but to continue to hold the stylus) would allow for microinteractions. There are just a few possible microgestures while holding a stylus but these are quite easy to perform and require low recognition effort.

5.4 Dual-Task Suitability

In summary, several parameters have an effect on how well two tasks suit a dual-task scenario, such as the duration of both tasks and the attention (alternate versus in parallel effort) that is necessary to solve the tasks without increasing the efficiency or //decreasing the?//effectiveness of the task. The suitability of two tasks depends on the level of required precision and the required attention as well as on the synchrony of these requirements.

Comparing the evaluated tasks, we argue that primary tasks, which have a long duration, are performed automatically and require low attention and motor effort, are suitable for simultaneous microinteractions. Of the conditions we evaluated, the palm grasp is the most promising for leaving enough motor resources for simultaneous hand gestures.

6 Conclusion and Design Guideline

Gestural interfaces lack the affordances and constraints that are readily provided by other interfaces, such as graphical and tangible ones [10, 11]. In particular, it is difficult to inform users what they are able to do, what they are currently doing or what they have just done. Because of this, gestural interfaces and in particular microgestures are not to be understood as a replacement for other kinds of interfaces, but rather as enabling novel ways of interaction. There are still many open questions to be answered, especially regarding the interaction opportunities and feedback representation.

Our taxonomy mainly investigated ergonomic interaction opportunities of microinteractions and can be used as a basis for designing microinteraction techniques for manual dual-task scenarios: First, the scenario has to be analyzed for defining the limits and requirements for microgestures. A gesture set can then be defined by looking at the formal structure of the chosen gestures. Lastly, a gesture driven decision about the sensory and tracking requirements of the hardware can be made.

6.1 Dual-Tasking Design

For a formal scenario design, we proposed two synergetic strategies: the economics of attentional and motor budgeting.

The selection of the primary and the secondary task is reasoned by the usage of different attentional resources. Our primary and secondary tasks used equal modalities by requiring tactile bio-feedback and kinesthetic self-awareness. An automatically performed primary task requires low attention [20]. This allows paying attention for simultaneously secondary tasks performance, such as microinteractions. These circumstances allow the economics of two tasks performances in parallel (Tab. 2, column 4). The example of steering a car, if it is done by people with some practice, represents an automatically performed task with low attention. Controlling the navigation system by microgestures could be a secondary one that requires attention.

The primary task defines the usage of motor resources as well as free potentials and available hand motor skills that can be used for simultaneous tasks. The grasp type that is performing the primary task (palm, pad, side) defines the motor resources which are used in the primary task (see Table 1, column 3). Our taxonomy identifies

microinteractions executable in parallel based on free motor resources (Table 2, column 1-3) and allows the creation of microgesture set for commanding the secondary task.

6.2 Interface Design

The developed gesture set defines requirements necessary for the interface design and the gesture tracking technique of microinteractions. For example tap-interactions should be tracked by a technology that provides a sequence of movement data like accelerometer. Gestures that are based on finger pressure are defined by vectored force applied on an object's or on skin surface and could be tracked by sensors that measures muscle activities such as EMG. The different tracking technologies shall be discussed for their data quality, and their interaction usability under different conditions given by both the microgesture design and the primary tasks.

There are some primary task-driven requirements for the sensor selections beside selecting the best suited sensors to measure formal gesture parameters. Covering the finger tips with interface components such as touch sensors would limit the tactile feedback (sense of touch) of the finger and the ability to conduct highly precise tasks. Moreover, the size and placement of the hardware could affect both the primary task and the ability to perform the input gestures. The interface design should not be cumbersome to wear and should be as small and unobtrusive as possible.

7 Further Research

The developed taxonomy serves as an analytic basis for systematic microinteraction design. As a next step, we intend to ask users to perform these microinteractions while performing a primary task and ask them to rate the feasibility of the gesture as well as scenario-related usability.

So far, we increased the hand gestures regarding their ergonomic structure and did not analyze their semiotic potentials. But within our interviews, we also received suggestions on what the gestures could communicate. For instance, the thumb-up-gesture (see Table 2, gesture 13) was commented to suit for okay-commands like answering the phone or selecting a pointed menu item. The taxonomy contains some more meaningful gestures, such as forming the index finger and the thumb to an "O" for communicating an "Okay". A snip gesture (see Table 2, gesture 12) could mean cutting something, and to put the thumb up (see Table 2, gesture 13) is also commonly understood as "Okay". When the gestures are linked to specific meanings and commands, it will be necessary to not just pay attention to the feasibility of a gesture but also to its potentials of association, guessability, and meaning.

References

1. Ashbrook, D.: Enabling Mobile Microinteractions, Doctoral Theses, Georgia Institute of Technology (2010)
2. Chewar, C.M., McCrickard, D.S., Ndiwalana, A., North, C., Pryor, J., Tesselndorf, D.: Secondary task display attributes: optimizing visualizations for cognitive task suitability and interference avoidance. In: Proc. Data Visualisation, pp. 165–171 (2002)

3. Czerwinski, M., Horvitz, E., Wilhite, S.: A Diary, Study of Task Switching and Interruptions. In: Proc. Conference on Human Factors in Computing Systems, pp. 175–182 (2004)
4. Feix, T., et al.: Grasp Taxonomy Comparison Sheet, http://web.student.tuwien.ac.at/~e0227312/documents/taxonomy_comparison.pdf
5. Harrison, C., et al.: Skinput: Appropriating the Body as an Input Surface. In: Proc. CHI 2010 (2010)
6. Howard, B., Howard, S.: Lightglove: Wrist-Worn Virtual Typing and Pointing. In: Proc. ISWC 2001 (2001)
7. Karam, M.: A Study on the Use of Semaphoric Gestures to Support Secondary Task Interactions. In: Proc. UIST 2001 (2003)
8. Lochair, C., Gustafson, S., Baudisch, P.: PinchWatch: A Wearable Device for One-Handed Microinteractions. In: Proc. MobileHCI 2010 (2010)
9. McCrickard, D.S., Chewar, C.M., Somervell, J.P., Ndiwalana, A.: A model for notification systems evaluation—assessing user goals for multitasking activity. *ACM Transactions on Computer-Human Interaction (TOCHI)* 10(4), 312–228
10. Norman, D.A.: Natural user interfaces are not natural. *Interactions* 17(3) (May-June 2010)
11. Norman, D.A.: Gestural Interfaces. A Step backwards in Usability. *Interactions* 17(5) (September-October 2010)
12. Oulasvirta, A., Tamminen, S., Roto, V., Kuorelahti, J.: Interaction in 4-Second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI. In: Proc. CHI 2005 (2005)
13. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K.E., Ansari, R.: Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)* 9(3), 171–193
14. Rekimoto, J., et al.: GestureWrist and GesturePad: Unobtrusive Wearable Interaction Devices. In: Proc. ISWC 2001, pp. 21–27 (2001)
15. Saponas, T., et al.: Enabling Always-Available Input with Muscle-Computer Interfaces. In: Proc. UIST 2009 (2009)
16. Spalteholz, W., Spanner, R.: *Handatlas der Anatomie des Menschen – Erster Teil: Bewegungsapparat*, Amsterdam, p. 284 (1960)
17. Tan, D., Morris, D., Saponas, T.S.: Interfaces on the Go, In XRDS. Crossroads. *The ACM Magazine for Students*, 30, doi:10.1145/1764848.1764856
18. Vardy, A., et al.: The WristCam as Input Device. In: Proc. ISWC 1999, pp. 199–202 (1999)
19. Wexelblat, A.: Research Challenges in Gestures: Open issues and unsolved problems. In: Proc. International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction 1997, pp. 1–11 (1997)
20. Wickens, C.D.: Processing resources in attention. In: Parasuraman, R., Davies, D.R. (eds.) *Varieties of Attention*, pp. 63–102. Academic Press, New York (1984)
21. Wolf, K, Dicke, C., Grasset, R.: Touching the Void: Gestures for Auditory Interfaces. In: Proc. TEI (2010)

Unifying Events from Multiple Devices for Interpreting User Intentions through Natural Gestures

Pablo Llinás, Manuel García-Herranz, Pablo A. Haya, and Germán Montoro

Dept. Ingeniería Informática, Universidad Autónoma de Madrid
C. Fco. Tomás y Valiente, 11, 28049 Madrid, Spain
{Pablo.Llinas, Manuel.GarciaHerranz, Pablo.Haya,
German.Montoro}@uam.es

Abstract. As technology evolves (e.g. 3D cameras, accelerometers, multitouch surfaces, etc.) new gestural interaction methods are becoming part of the everyday use of computational devices. This trend forces practitioners to develop applications for each interaction method individually. This paper tackles the problem of interpreting gestures in a multiple ways of interaction scenario, by focusing on the abstract gesture rather than on the technology or technologies used to generate it. This article describes the Flash Library for Interpreting Natural Gestures (FLING), a framework for developing multi-gestural applications integrated and running in different gestural-platforms. By offering an architecture for the integration and unification of different types of interaction, FLING eases scalability while presenting an environment for rapid prototyping by novice multi-gestural programmers. Throughout the article we analyse the benefits of this approach, comparing it with state of the art technologies, describe the framework architecture, and present several examples of applications and experiences of use.

Keywords: FLING framework, Multi-touch interface, multiple input peripherals, application development.

1 Introduction

The evolving progression of HCI interfaces promises a brilliant future to seamless control and handling of intelligent devices. New ways of interaction are invented and integrated in the search for the best and most natural method of interaction between humans and computers. Their goal is to allow for a more intuitive and natural way of expression when communicating with computers. An immediate consequence of this evolution is the creation of new input devices with which we can operate. Today, intelligent homes can be fitted with all sorts of sensors (temperature, movement, pressure, fingerprint...). Also, existing devices are enhanced with extra sensing capabilities [1], such as multi-touch screens, accelerometers and voice recognition in smartphones like Apple's iPhone or Google's Nexus S.

As computers integrate deeper into our daily lives, the most natural way to use them becomes dependent to each particular scenario. A single application will have to allow different interaction methods across different settings in order to offer the most suitable experience each time. For example, for talking to a friend, video conferencing

will be suitable at home, where you can give visual and hearing attention to the other person, while text chatting will be more adequate in a football stadium during a match, where being heard can be very difficult.

From a programmer's point of view, more input devices entail a higher complexity when treating input. Different peripherals offer different interaction possibilities, which generate input events that need to be processed in order to interact with the applications. To reduce the difficulty of dealing with a large (and sometimes unknown) number of input devices, cross-platform frameworks are used to take care of these devices, and for developing applications using a set of abstract input events.

The difficulty of programming for cross-platform ubiquitous control has been shifted from treating each input device, to learning how to work with cross-platform frameworks. In a world of constant upgrades in hardware gadgets and increasing intelligent sensing devices, developers need unification and standardization of user intention recognition regardless of the technology employed.

2 Related Work

The advantages of allowing the user to choose the most adequate input device for interaction with applications are therefore being studied and proven profitable. Multimodal interaction frameworks [2, 3, 4] pose valuable precedents of frameworks for using different input devices for human-computer interaction, using input events either independently or combined. However, we lack a higher degree of flexibility when translating device actions (fixed for each input peripheral) into application actions (which trigger the available operations on the computer). One programmer could require a sequence of device events to carry out an operation, while another would be looking for a different one. Moreover, different users using the same application may require different responses to similar input patterns.

This issue is of particular importance in multitouch environments in which events are usually composed of several device inputs (e.g. multiple fingers touching a display) that have to be, furthermore, interpreted from complex raw data, such as blob identification and analysis. In order to ease the process of interpreting device inputs in camera-based multitouch systems, several low-level frameworks such as TouchLib [5], reactIVision [6], Community Core Vision (CCV) [7] or Touchè [8] are available to identify significant blobs as screen touches ready to process. This interpretation results, in most cases, in TUIO messages codifying each finger and its evolution on the surface. This kind of frameworks provides a first level of abstraction, allowing applications to listen from a single channel to every finger interaction in a unified manner. Advanced examples of these are BBTouch [9] and LightTracker [10], which improve the aforementioned frameworks by allowing an advance tuning of the recognition parameters, or VVVV [11], which allows associating blob input with different visualization methods using a visual programming paradigm.

Nevertheless, multitouch interactions are most of the times composed of several fingers and, therefore, multiple events have to be reinterpreted, according to the element of the UI over which they are acting, to form a high level global gesture (e.g. two fingers moving are interpreted as moving apart from each other). Finally, the interpreted gestures have to be associated with a particular action. Some systems such as PyMT [12] or Grafiti [13] provide a transparent mechanism to distribute events

among the elements of the UI as well as basic interpreters for the most common gestures such as move, resize or rotate.

However, the number of different gestures in multitouch systems can grow far beyond the basic ones and the actions associated with each of them may vary among applications or, in a single application, from component to component. Thus, a higher degree of abstraction, modularization and composition is needed. Systems such as Surface SDK [14] or DiamondTouch SDK [15] provide this kind of flexibility to a reasonable degree but are constrained to a particular platform, limiting their extensibility and scalability in an ever-increasing world of multitouch hardware solutions. Similarly, proprietary frameworks such as GestureWorks [16] constrain their extensibility, compared to their open source counterparts, in an ever-increasing world of gestures.

Nonetheless, while multitouch systems are gaining popularity, they are far from being an alternative to the standard mouse-keyboard paradigm and will have to further cohabitate with new interaction devices such as the Wii remote or Kinect. Thus, systems such as Squidy [17] provide a low-level alternative to unify various device drivers, frameworks, and tracking toolkits in one common library, overcoming the limitations of higher level solutions such as GestureWorks [16] or Grafiti [13] designed to work just with multitouch events.

From the high level open source alternatives designed to support different input mechanisms, define new gestures and dynamically associate actions with them, we can distinguish between language dependent and language independent frameworks. Those that rely on a separated event system, allowing to program in the language of our choice, force the programmer to provide a description of the UI and its components to the event interpreter layer. This is done either explicitly, therefore, adding an extra complexity to the programming process that prevents a rapid prototyping, as in SparshUI [18] and Midas [19], or implicitly through a widget library, as in libTISCH [20], simplifying the communication channel but making very difficult to interpret events outside the boundaries of the widget.

MT4j [21] (MultiTouch for Java) is, on the other hand, a language dependent framework designed for rapid prototyping and gesture extension. Making use of the well-known event-listener java architecture, it allows components to listen to particular gestures without modifying or adding complexity to the UI design and coding.

FLING, the framework described in this article, falls into this last category but relies on Adobe's Flash, instead, as a well-known platform to graphical designers. Thus, graphic design and program logic can be easily separated and distributed among designers and programmers, allowing for good-looking rapid prototyping. In addition, contrary to systems such as MT4j, FLING provides a double distribution mechanism. Through one channel, interpreted gestures are propagated to every component, allowing them to know what is happening to themselves as well as to the rest of the components. Through the second one, raw events are propagated too, allowing to program global or partial interpretations in any component of the UI, whether they fall in or outside its boundaries.

3 The FLING Framework

FLING (Flash Library for Interpreting Natural Gestures) is a cross-platform multi-gesture framework for developing Adobe Air and Flash applications using the ActionScript 3.0 programming language.

FLING has been developed under the following principles:

3.1 Platform-Independent

FLING shares the “write once, run anywhere” philosophy. In order to run on the highest number of computing devices, a platform-independent programming language is a must. ActionScript 3.0, the language behind Adobe Flash and Air applications, was chosen for this purpose. Applications run on any desktop operating system (Windows, Mac and Linux) and also on Android smartphones. A special version, Flash Lite, can even be used on more basic mobile phones [22].

We chose this way of deploying cross-platform applications instead of using adapted interfaces because we believe that technologies such as Java, HTML and Adobe Flash, which in the past have been secluded to being used on standard PCs, will soon be running with the same capabilities on even the most basic portable devices. The processing gap between different computing architectures is closing-in, and today we can find smartphones¹, tablets², notebooks³ and desktop computers sharing very similar dual-core processing computational power.

The Adobe family of products provides a very powerful and robust set of tools for the visual design of graphical interfaces. Also, when integrating these graphical elements into the Flash platform for adding programming logic, there is full compatibility and interoperability between multimedia contents. A drawing made in Photoshop can be imported into Illustrator to get a vector graphic that then can be inserted into Flash and get animated. The final symbol can be accessed and manipulated from code using the ActionScript language. And finally, the result will be an Adobe Air or Flash multi-platform application.

3.2 Useful for Rapid Prototyping

For first-time cross-platform multi-gesture application developers, FLING provides a basic manipulation of visual objects using the traditional mouse and keyboard, and multi-touch surfaces. Extending from the base FLING object class, object movement, rotation, resizing and physics engine (inertia, collisions, gravity...) can be enabled with a single line of code, as seen below.

Example of extending from the base FLING object class

```
public class NewObject extends FlingObj{
    public function NewObject():void{
        movable = resizable = rotatable = physics = true;
    }
}
```

The visual symbol of class “NewObject” will respond according to the activated capabilities upon standard input events. In the case of the mouse and keyboard, the

¹ http://www.pcworld.com/article/204947/lg_announces_smartphones_with_dualcore_processor.html

² <http://www.tgdaily.com/mobility-features/49854-nvidia-showcases-dual-core-tegra-2-tablet>

³ <http://liliputing.com/2010/10/samsung-launches-nf310-dual-core-netbook-with-hd-display.html>

object will respond as shown on Table 1. Using a multi-touch surface, the object will respond as shown on Table 2.

This initial functionality allows for basic application interaction without getting into event interpretation or gesture handling. It is oriented towards Adobe Flash and Air developers with minimal knowledge in device input processing, and allows them to create cross-platform multi-gesture applications which can run on multi-touch surfaces and standard computers without making the effort of learning a complex new framework.

FLING requires little adaptation for first-time programmers. One of the basic fundamentals of its operation is a tree-structured hierarchy of symbols or visual objects (Figure 1). A unique FLING root object, representing the application itself, contains all device parsers and gesture interpreters. All symbols used in an application must be children to this root object (either directly, or further down in the tree of child objects). This is required because it represents the order in which visual objects are layered in an application. The root object is the background of the application, and it represents the lower-most layer. The next level in the hierarchy of children objects represents the layer on-top of the background layer, and so on.

Table 1. Default actions triggered by mouse and keyboard input events

Input device gesture	Action performed on object
Drag with mouse	Object moves under mouse pointer
Control key + Drag with mouse	Object resizes in regard to its center and the mouse pointer
Shift key + Drag with mouse	Object rotates in regard to its center and the mouse pointer

Table 2. Default actions triggered by multi-touch finger input events

Input device gesture	Action performed on object
Slide one finger over the object	Object moves according to the finger movement
Pose two fingers over object and separate or join them	Object resizes according to distance variation between fingers
Pose two fingers over object and move one around the other	Object rotates according to angle variation of line joining fingers

The FLING framework relies on Flash's native visual hierarchy of objects for target identification, allowing disambiguating when objects overlap. As all Flash objects have one (and only one) parent and their insertion order decides among siblings, no inconsistencies can occur between FLING's object targeting and Flash's visual representation. Using the native structure for object nesting presents a logical way to navigate through objects and can, therefore, be already found in most applications.

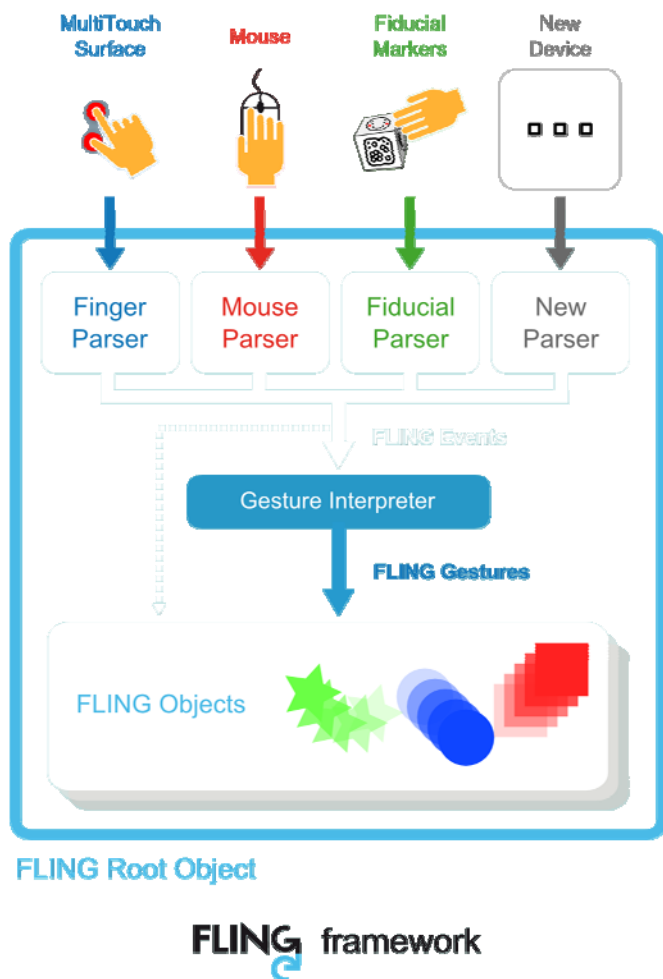


Fig. 1. Event parsing and interpretation inside the FLING framework

3.3 Allows Customization of Event Interpretation and Triggered Actions

To extend the default functionality offered as standard when working with FLING, developers can tune and enhance the interpretation of device events into gestures, and the reactions upon receiving interpreted gestures. This can be achieved with minimal code modification because of the predisposition to customization.

The three main elements with which FLING works are events, gestures and actions. Events are signals from input devices that are processed and homogenized into a common stream of events by the device parsers. Each device has an associated event parser that in some cases makes use of external drivers to capture input signals.

Gestures are interpreted by another module which receives events parsed from all the input devices available to the user. The interpreter can use the combination of

these input events and knows of all existing interface objects in order to make sense of the user's intentions. It outputs FLING Gestures, which consist of recognized user intentions. FLING is then responsible of propagating and making gesture events reach their correct target. Once an interactive object receives a gesture event (or FLING Gesture), it reacts according to the triggered action for that gesture. Some triggered actions are preconfigured by default (as is the case of the move, rotate and resize actions mentioned in section 3.2) and others are left blank, but all of them are customizable to fit the needs of each application.

As an example of triggered action customization, we will take the resizing gesture and change its default action so that instead of changing size, the target object will change its transparency. The code added to the target object for modifying this behaviour is as follows:

Example of triggered action customization

```
override public function onRotateGesture(gesture:FlingGesture):void{
    this.alpha += gesture.varAngle%360;
}
```

Normally, the “varAngle” property of the FLING event received is used to rotate the object accordingly. Instead, we are using this value, normalized between -1.0 and 1.0, to alter the alpha value of the object, hence changing its transparency. This is just an example of how easy it is to change the behavior associated to a particular gesture.

Another example of framework customization can be seen when modifying the way device events are interpreted into gestures. We will take the multi-touch gesture interpreter and add some code to recognize a new gesture. The new gesture will consist in a quick slide of two fingers (*cursor1* and *cursor2*) over a visual object, from top to bottom. This gesture will be called the “minimize” gesture, and could be linked to the minimizing action, but this is entirely up to the programmer.

The code needed inside the finger interpreter to recognize this gesture is the following:

Example of gesture customization

```
if(numCursors == 2){
    if(cursor1.type == cursor2.type == CURSOR-EXIT){
        if(slideDirection(cursorStream1)
           == slideDirection(cursorStream2) == SLIDE-DOWN){
            flingGestures.push(new FlingGesture("MINIMIZE"));
        }
    }
}
```

3.4 Cross-Platform Multi-gestures

Another requirement for ubiquitous control is the possibility of multiple device input (Figure 2). FLING comes with parser modules for mouse & keyboard, multi-touch surfaces (such as Microsoft's Surface⁴, MultiTouch Cell⁵ and the reactTable [23]),

⁴ <http://www.microsoft.com/surface/>

⁵ <http://multitouch.fi/products/cell/>

pressure tokens (special objects recognizable by pressure marks) and fiducial markers [24]. These input devices can be used right away, and can be configured and customized to fit the application's needs.

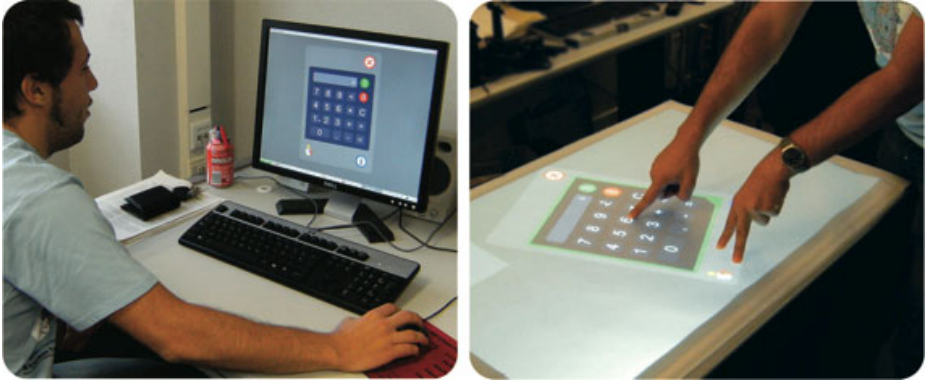


Fig. 2. A multi-platform running environment and multi-gesture input interaction enable applications run on the device which most suits each scenario

For advanced necessities, it also allows for new input devices to be used when needed. The steps for adding a new input device are:

1. Identify the input data format in which the device sends events from the user's interaction. If not provided by the standard Adobe Flash / Air libraries, parse the device signals into atomic non-interpreted events. These atomic events should match the device's interaction possibilities. For example, a joystick will have coordinate events indicating its position, and button events indicating any change in the state of their manipulation.
2. Add a gesture interpretation module for the new input device, which groups atomic input events into recognized gestures. For example, using multi-touch surfaces, two fingers separating from each-other over a same object are recognized as a resize event by default. Gesture metrics (e.g.: resizing value) are also included in the interpreted gesture. The gesture interpretation module sends recognized gestures, called FLING events, as shown on Figure 1, which can be equivalent to those already handled, or new ones which are exclusive to each input device.
3. FLING will forward both atomic (or raw) events and gesture (or FLING) events to all objects, as described in section 3.2, for them to react consequently. All of the aspects of input device reading and interpreting can be openly customized to obtain the required functionality, as stated on section 3.3.

Multi-touch surfaces are one of the input devices currently supported by FLING (Figure 3). On a typical multi-touch table, fingers placed on-top of the touch panel (A) create light blobs that can be tracked by a video camera [25]. The camera sends a video stream (1) to a blob driver (B), such as reactIVision [6], Touchlib [5] or Community Core Vision [7], which then outputs finger events using the TUIO

protocol [26] through a data socket (2). FLING (C) connects to this data socket and uses the incoming finger events from the panel to begin the interpretation and propagation processes (3). Finally, the application reacts to the intentions expressed by the user, and evolves its visual interface (4) which is displayed on the same surface used as touch panel (D).

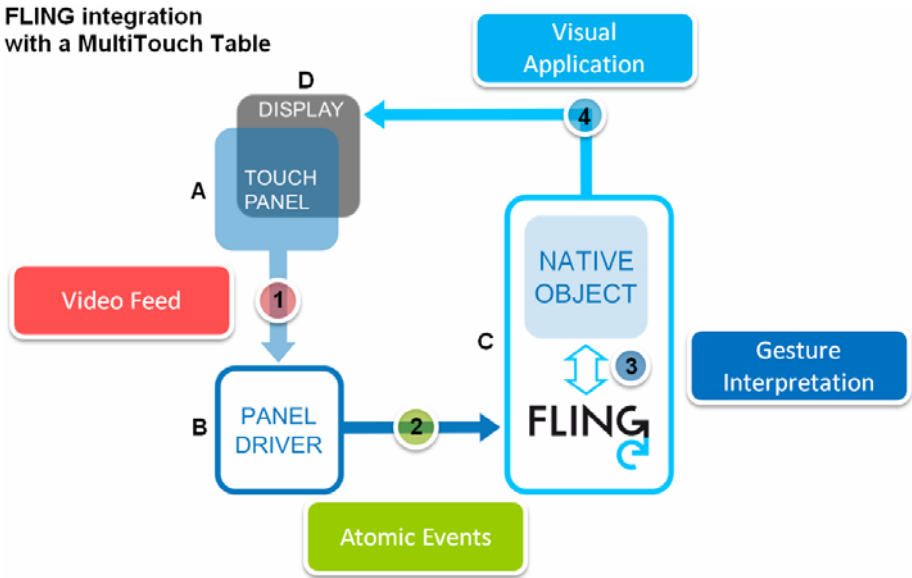


Fig. 3. Multi-touch surface input handling using the FLING framework

Inside the FLING framework, a parser module reads unprocessed finger events obtained from the hardware drivers, and homogenizes them into standard input device events. This homogenization consists in a translation of native signals to a common event class, which holds the same information but in a standard form. Then, the finger interpretation module is in charge of interpreting and generating gesture events that FLING will send to the appropriate application object. Interactive objects will react in response to these gesture events, and will also receive the unprocessed input events in case more information is required.

The format of gestures produced from a multi-touch surface is compatible with those generated using mouse, fiducial or token devices. Interactive objects from the application receive complete and descriptive gestures with the user’s intentions, but don’t have to worry about the input used to express those intentions. Additionally, unprocessed events from the input devices are also delivered to application objects for the event of needing detailed information about raw input data.

3.5 Progressive Learning Curve

The idea behind offering both rapid prototyping and advanced customization capabilities is to create a progressive and smooth learning curve when working with

the framework. Existing cross-platform frameworks offer extensive functionality and very advanced user expression recognition, but are difficult to use at first and require a lengthy learning period. We experienced these difficulties when trying commercial products such as Gestureworks[16] or the TUIO Flash client library [27] for receiving multi-touch events. It is normal to go through a number of learning steps while getting comfortable using a new framework, and a learning curve similar to the one described by Gaines [28] is typically experienced.

4 Sample Applications

FLING has been thoroughly tested and used for the development of full applications which have made their way into educational and experimental projects. Different working environments with adapted input devices have accommodated these applications, and people with varying levels of knowledge and expertise have given them a try.

4.1 Therapy Applications

To serve as a complement for people with disorders, games such as Simon® and Gesture Hero (Figure 4) were developed under particular requirements. They were designed in cooperation with teachers and assistants of people with Down syndrome and Alzheimer’s disease to serve as therapy work for these collectives.

This supervised use has served to gather information about interaction habits and difficulties that has been used in FLING’s design.

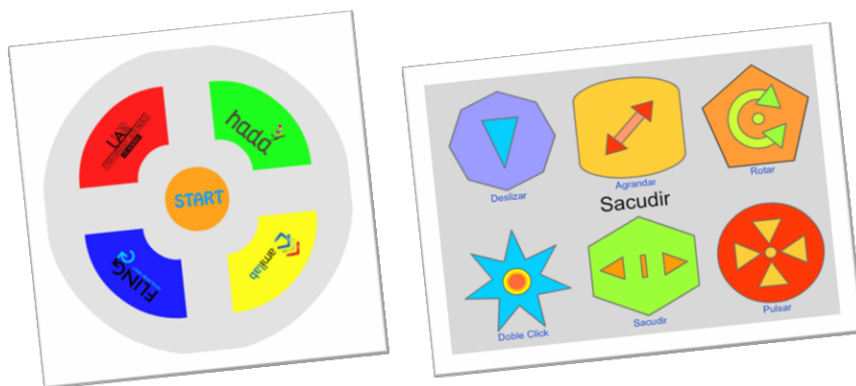


Fig. 4. On the left, the Simon® game, a memory training application. On the right, the Gesture Hero game, a psychomotor skill game.

Specific gestures were added to the Gesture Interpreter module in order to supply these applications of their interaction requirements.

4.2 Educational Applications

In addition to the therapy applications, other educational applications have been developed to be included in the regular activities of people with Down syndrome during their classrooms. Examples are the Postman Game and the Price is Right game shown on Figure 5.

The Postman Game consists in handing packages to the right recipient using the same procedure as real postmen do in our university. The user needs to search for the correct floor, office or desk assigned to the recipient from an address file. Then she must drag the package on to the correct mailbox. This application was made to train future university postmen, and it resembles the real procedure faithfully.

The Price is Right game simulates a typical trip to the cafeteria. The waiter offers the user a product at random, and informs her of its price. In the wallet the user has money represented by actual photographs of real Euro bills. She must drag money from the wallet onto the waiter's plate, and the waiter will give her the right change in return. Upon analysing the payment, the game awards the user with a rating which will be better as the payment gets closer to the right price. The change from the previous purchase is used to continue throughout the game.



Fig. 5. On the left, the Postman Game, a package delivery simulator. On the right, the Price is Right, a money management game.

These games are oriented to desktop computer usage, although they have also been tested on multi-touch tables with real-case users. This was possible thanks to the variety of input devices supported by FLING, which made it easy to shift from one platform to another without any modification.

4.3 Data Visualization and Control Applications

One of the great advantages of using a development environment as Adobe Air / Flash is its inherited multimedia capabilities. Video and audio content can be accessed in many ways, and FLING adds a rich interaction experience that helps in the complex task of data visualization of large collections of multimedia content.

For these reasons, FLING was employed at Carnegie Mellon University to create a video discovery application that used a geographical information system based on Google Maps™. Maps can be browsed with multi-touch gestures, and embedded videos can be opened and played with full timeline controls, which are easily triggered with finger gestures. The typical hardware setup for this application is a vertical multi-touch panel on which the graphical interface can be projected and the interaction is executed using the hands. This application is an example of different interaction techniques seamlessly combined through FLING. While the Google Maps™ API listens to mouse events, the multitouch panel generates TUIO events. In addition, the rest of the elements of the interface, such as the opened videos, listen to finger events. FLING integrates all the interactions seamlessly, significantly reducing the programmer's effort. Google Maps™ are integrated naturally, as they do in traditional PC applications, and all the functions of their API can be used directly out of the box.



Fig. 6. On the left, a video discovery application using GIS developed at the Instinctive Computing Lab on Carnegie Mellon University. On the right, a control application for an intelligent room at the Universidad Autónoma de Madrid.

The intelligent environment control application is meant to be run on a multi-touch table that accepts fiducial marker interaction. These fiducial symbols can be placed over appliances drawn on the house map to control some of their properties. For example, the intensity fiducial marker can be placed over lamps, and can be rotated clockwise to increase light brightness, or anti-clockwise to reduce it. A selection fiducial can be placed over the TV to change the channel. This application was developed to show how new input devices (fiducials in this case) can be easily added to FLING, and how its functionality is automatically incorporated to the existing gestures.

5 Conclusions and Future Work

This paper has focused on the problem of developing applications for an increasing number of interaction mechanisms. In doing so, we have stressed the necessity to distinguish between events, gestures and actions. Events depend on the interaction

mechanisms, gestures represent the users' manipulation intentions and actions depend on each particular application.

Following this distinction we have developed FLING, a framework that allows:

- To easily build interaction independent applications without having to deal with interaction events.
- To build new complex gestures combining events from multiple interaction sources.

This framework is designed to be:

- Scalable: so that new interaction mechanisms and gestures can be added and integrated with existing ones.
- Multi-platform: so that it can run in the varied number of platforms in which the new interaction mechanisms are emerging.

This framework has been tested, seamlessly and over a number of platforms, through a number of applications using both traditional mouse and keyboard, as well as novel multi-touch interaction mechanisms.

This has been achieved through a modular design, separating the parsing of each input device from the gesture interpreter process. Thus, new sensors can be added by just incorporating the corresponding parser to the framework.

We are currently working in extending the interaction experience through new interaction mechanisms and merging the interaction of multiple interfaces.

Finally, working in the Ambient Intelligence domain, we look forward to enriching the interactions with context-aware information in a multi-user, distributed, intelligent environment.

Acknowledgments. This work has been partially funded by the following projects: ASIENS: Adapting Social & Intelligent Environments to Support people with special needs (Ministerio de Ciencia y Educación de España, TIN2010-17344), and Vesta (Ministerio de Industria, Turismo y Comercio de España, TSI-020100-2009-828).

References

1. Lester, J., Hurvitz, P., Chaudhri, R., Hartung, C., Borriello, G.: MobileSense-Sensing modes of transportation in studies of the built environment. In: *UrbanSense 2008*, pp. 46–50 (2008)
2. Dragicevic, P., Fekete, J.: Input device selection and interaction configuration with ICON. In: Blanford, A., Vanderdonk, J., Gray, P. (eds.) *People and Computers XV Interaction without Frontiers: Joint Proceedings of IHM 2001 and HCI 2001 (IHM-HCI 2001)*, pp. 543–558. Springer, Heidelberg (2001)
3. Flippo, F., Krebs, A., Marsic, I.: A framework for rapid development of multimodal interfaces. In: *5th International Conference on Multimodal Interfaces (ICMI 2003)*, pp. 109–116. ACM, New York (2003)

4. Serrano, M., Nigay, L., Lawson, J., Ramsay, A., Murray-Smith, R., Deneff, S.: The openinterface framework: a tool for multimodal interaction. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems (CHI EA 2008), pp. 3501–3506. ACM, New York (2008)
5. Touchlib: an opensource multi-touch framework, <http://www.whitenoiseaudio.com/touchlib>
6. Kaltenbrunner, M., Bencina, R.: reacTIVision: a computer-vision framework for table-based tangible interaction. In: 1st International Conference on Tangible and Embedded Interaction (TEI 2007), pp. 69–74. ACM, New York (2007)
7. Community Core Vision, <http://ccv.nuigroup.com/>
8. Touchè, <http://gkaindl.com/software/touche>
9. Bederson, B.B., Grosjean, J., Meyer, J.: Toolkit Design for Interactive Structured Graphics. *IEEE Trans. Softw. Eng.* 30(8), 535–546 (2004)
10. Gokcezade, A., Leitner, J., Haller, M.: LightTracker: An Open-Source Multitouch Toolkit. *J. Comput. Entertain.* 8, article 19 (2010)
11. VVVV, <http://vvvv.org/>
12. Hansen, T.E., Hourcade, J.P., Virbel, M., Patali, S., Serra, T.: PyMT: a post-WIMP multi-touch user interface toolkit. In: ACM International Conference on Interactive Tabletops and Surfaces (ITS 2009), pp. 17–24. ACM, New York (2009)
13. De Nardi, A.: Graffiti: Gesture Recognition mAnagement Framework for Interactive Tabletop Interfaces. Diploma thesis. University of Pisa (2008)
14. Surface SDK, <http://msdn.microsoft.com/en-us/library/ee804845.aspx>
15. Esenther, A., Forlines, C., Ryall, K., Shipman, S.: DiamondTouch SDK: Support for Multi-User, Multi-Touch Applications. Mitsubishi Electronics Research Laboratory, Report No. TF2002-48 (2002)
16. Gestureworks, <http://gestureworks.com/>
17. König, W.A., Rädle, R., Reiterer, H.: Squidy: a zoomable design environment for natural user interfaces. In: 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA 2009), pp. 4561–4566. ACM, New York (2009)
18. Ramanahally, P., Gilbert, S., Niedzielski, T., Velázquez, D., Anagnost, C.: Sparsh UI: A Multi-Touch Framework for Collaboration and Modular Gesture Recognition. In: Proc. of WINVR 2009, Conference on Innovative Virtual Reality, pp. 1–6 (2009)
19. Scholliers, C., Hoste, L., Signer, B., De Meuter, W.: Midas: a declarative multi-touch interaction framework. In: 5th International Conference on Tangible, Embedded, and embodied Interaction (TEI 2011), pp. 49–56. ACM, New York (2011)
20. Echtler, F., Klinker, G.: A Multitouch Software Architecture. In: 5th Nordic Conference on Human-Computer Interaction (NordiCHI 2008), pp. 463–466 (2008)
21. Laufs, U., Ruff, C., Zibuschka, J.: MT4j - A Cross-platform Multi-touch Development Framework. In: Engineering Patterns for Multi-Touch Interfaces 2010, Workshop of the ACM EICS (2010)
22. Blom, S., Book, M., Gruhn, V., Hrushchak, R., Kohler, A.: Write Once, Run Anywhere A Survey of Mobile Runtime Environments. In: 3rd International Conference on Grid and Pervasive Computing – Workshops, pp. 132–137. IEEE Press, New York (2008)
23. Jordà, S., Geiger, G., Alonso, M., Kaltenbrunner, M.: The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces. In: 1st International Conference on Tangible and Embedded Interaction (TEI 2007), pp. 139–146. ACM, New York (2007)

24. Bencina, R., Kaltenbrunner, M.: The design and evolution of fiducials for the reactivation system. In: 3rd International Conference on Generative Systems in the Electronic Arts (3rd Iteration 2005), Melbourne, Australia (2005)
25. Wang, F., Ren, X., Liu, Z.: A Robust Blob Recognition and Tracking Method in Vision-Based Multi-touch Technique. In: International Symposium on Parallel and Distributed Processing with Applications (ISPA 2008), pp. 971–974. IEEE Press, New York (2008)
26. Kaltenbrunner, M., Bovermann, T., Bencina, R., Costanza, E.: TUIO: A protocol for table-top tangible user interfaces. In: 6th Int'l. Workshop on Gesture in Human-Computer Interaction and Simulation (2005)
27. TUIO Flash client library, <http://www.tuio.org/?flash>
28. Gaines, B., Shaw, M.: A learning model for forecasting the future of information technology. *J. Future Computing Systems* 1, 31–69 (1986)

SimpleFlow: Enhancing Gestural Interaction with Gesture Prediction, Abbreviation and Autocompletion

Mike Bennett^{1,2}, Kevin McCarthy², Sile O'Modhrain³, and Barry Smyth²

¹ SCIEN, Department Of Psychology, Stanford University

² School of Computer Science, University College Dublin, Ireland

³ Sonic Arts Research Centre, Queens University, Belfast, UK

mikemb@stanford.edu

Abstract. Gestural interfaces are now a familiar mode of user interaction and gestural input is an important part of the way that users can interact with such interfaces. However, entering gestures accurately and efficiently can be challenging. In this paper we present two styles of visual gesture autocompletion for 2D predictive gesture entry. Both styles enable users to abbreviate gestures. We experimentally evaluate and compare both styles of visual autocompletion against each other and against non-predictive gesture entry. The best performing visual autocompletion is referred to as SimpleFlow. Our findings establish that users of SimpleFlow take significant advantage of gesture autocompletion by entering partial gestures rather than whole gestures. Compared to non-predictive gesture entry, users enter partial gestures that are 41% shorter than the complete gestures, while simultaneously improving the accuracy (+13%, from 68% to 81%) and speed (+10%) of their gesture input. The results provide insights into why SimpleFlow leads to significantly enhanced performance, while showing how predictive gestures with simple visual autocompletion impacts upon the gesture abbreviation, accuracy, speed and cognitive load of 2D predictive gesture entry.

1 Introduction

Gestural interfaces are now a familiar mode of user interaction and gestural input is an important part of the way that users can interact with such interfaces. Gestural input accommodates a style of interaction that goes beyond the point-and-click of traditional WIMP-based interfaces and allows users to interact in a more intuitive and efficient [28, 31, 9, 16]. For example, simple swiping motions (whether mouse pointer, trackpad, or direct screen-contact) can be used as more intuitive navigation controls, circumventing the precision that is demanded by more traditional point-and-click interactions [33]. A range of more sophisticated gestures, such as a two-fingered 'pinch' for zooming, are now an increasingly familiar part of gesture-based interaction dictionaries [13, 15].

Ultimately, gesture-based interaction promises to provide users with a more intuitive and richer interaction vocabulary, offering greater interaction bandwidth for lower input effort [33, 11]. However, this is not always the case and certainly the quest for more expressive gestures can lead to suboptimal gestures that are difficult

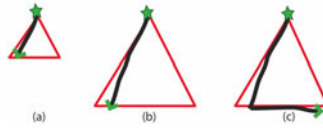


Fig. 1. Example of SimpleFlow. Shown is the visual feedback displayed when entering a triangle gesture. Star is starting point of gesture entry. Black line is the gesture entered so far by the user - (a) is time point 1, (b) time point 2, (c) time point 3. Red line is the real-time predicted gesture.

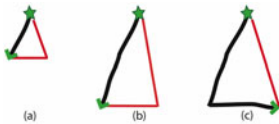


Fig. 2. Example of Scale Free Dynamic Paths shown when entering a triangle gesture. Star, black line & red line have the same meaning as in Figure 1.

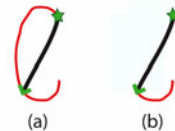


Fig. 3. Example of erroneous gesture predictions that could be shown when entering a triangle gesture. (a) SimpleFlow, (b) SF-Path.

for users to produce reliably, leading to interaction failures and frustration. One way to help users as they produce gestures is to provide autocompletion suggestions as the gesture unfolds [4, 30, 5, 13]. If this can be done efficiently and accurately then the user will benefit in a number of ways. For example, by accepting a gesture suggestion the user can complete their gesture early and so gestural input can be made more efficient [27, 11, 6]. Moreover, by encouraging early gesture completion it removes the risk of errors that might have occurred if the gesture had to be manually completed. At the same time there may be additional costs for the user, especially if they find the gesture completions to be difficult to accommodate into the work-flow; if completion suggestions are inaccurate, for example, then users will quickly become frustrated by such inconvenient interruptions.

In this paper we describe two approaches to gesture autocompletion (we refer to these as *predictive gestures*), both of which present the user with visual feedback as a gesture is entered (Figure 1 & 2). Both also provide the user with an option to auto-complete the target gesture early. In addition, we describe the results of a detailed user evaluation, in which the predictive gestures are compared to manual (non-predictive) gesture input, focusing on the gesture shortening, accuracy, speed and cognitive load characteristics of these techniques. In particular we demonstrate and explain how predictive gesture techniques can lead users to significantly abbreviate gestures while improving the speed and accuracy of their gesture input; along with users stating a clear preference for the predictive gesture input over non-predictive entry.

The contributions of this work are:

1. Based on the visual autocompletion styles in this paper, our results establish that users will shorten predictive gestures by a significant amount. For predictive text entry it is well established that users will significantly shorten words during text

entry. Previous to this it had not been established whether users of predictive gesture entry systems will enter shorter gestures, nor had it been established by how much they will shorten the gestures by.

2. Does allowing gesture abbreviation and showing visual autocompletion during gesture entry increase the cognitive demands on users? Unexpectedly, we find that the cognitive load does not increase for the autocompletion styles.
3. We introduce SimpleFlow, an effective style of visual autocompletion, which subtly but for users importantly differs from the simplest form of visual feedback (Scale Free Dynamic Paths).
4. Are there significant trade-offs between predictive versus non-predictive gesture entry and the speed and accuracy with which gestures are entered and recognised? We establish interactions and tradeoffs between gesture autocompletion, gesture abbreviation, input speed and input accuracy for SimpleFlow and Scale Free Dynamic Paths.

2 Related Work

Entering gestures accurately and quickly can be challenging [3, 32, 26]. Some of the challenges arise because of the need for algorithms that accurately recognise gestures [32, 25]. Other challenges include questions about what interface designs provide beneficial visual feedback before, during and after gesture entry [5, 3].

Visual Feedback For Predictions. Various styles of visual feedback have been proposed for gesture entry [13, 31, 5, 18, 2, 30, 33, 22]. Often the feedback styles are for enhancing pre- and post-gesture entry. Research on pre-gesture feedback aims to help users know and remember what set of gestures are available, while post-gesture feedback helps users understand whether they successfully entered a gesture, and if not what went wrong during gesture entry.

ShadowGuides [13], OctoPocus [5], Fluid Sketches [4] and others [18, 24, 2] are examples of interaction techniques which provide real-time visual feedback during gesture entry. These aim to help users learn the gestures better and help them understand how well they are entering the gestures. Other forms of real-time feedback help the user understand what the results of the gesture will be, such as telling them what actions will be performed, e.g. show the user where a dragged icon will end up [6], or what action will be performed [5].

Bau et al. [5] propose a very useful classification framework for the various styles of visual feedback. Within their classification framework SimpleFlow is a dynamic guide, with feedback that has a continuous update rate, a real recognition value, a gesture filter and a gesture representation.

Applying Gestures. Recently there has been a surge in applications of gestures without visual feedback and in more novel contexts, such as using human skin as a gesture input surface [16], or moving your hands freely in the air as a form of bimanual gesture entry [15].

Other styles of gesture input which are less obviously gesture prediction include Drag-and-Pop / Drag-and-Pick [6], Push-and-Pop [11], Escape [33], Dasher [31] and

marking menus [23]. In those examples the effects of users' physical actions are enhanced, so they can perform shorter and less actions than normally required for the tasks, which can be thought of as forms of gesture prediction and abbreviation.

Modeling Gesture Interaction. Models of human performance when following trajectories have also been investigated [1], and could prove useful for modeling gesture input [10, 20], as models and metrics of human performance have proven useful for predictive text input [27, 29]. Related to modeling gestures is research around measuring and modeling the complexity of gestures [26], which could be applied to quantifying the predictability of gestures.

3 Visual Autocompletion and Predictive Gestures

In this section we present¹ two styles of predictive visual feedback, which are shown during gesture entry. Both styles show visual predictions of the target gesture a user is inputting. In this way they allow the autocompletion of gestures and thus encourage quicker and more efficient gesture input. We also outline the non-predictive visual feedback commonly found in many gesture entry applications [12, 8, 14], which we experimentally compare the predictive visual feedback against.

3.1 Visual Autocompletion

We implemented two styles of visual autocompletion for predictive gestures. Both styles are closely related and subtly but importantly differ in how they provide the real-time visual feedback. The first style of visual autocompletion is called *SimpleFlow*, as shown in Figure 1, and the second style is called *Scale Free Dynamic Paths* (SF-Path), as shown in Figure 2.

The purpose of SimpleFlow and SF-Path is to keep the user informed during gesture input, such that they know as early as possible which of the trained gestures (examples in Figure 4) matches their input.

Both styles are designed to enhance user performance during gesture entry, by improving the speed and accuracy with which gestures are entered. Both are also designed to enable users to abbreviate gestures. The ability to enter abbreviated gestures resembles predictive text entry systems - where users can enter complete words by only typing the first few letters of the words [27, 29].

Figures 1 & 2 show examples of what users see on-screen while using SimpleFlow and SF-Path. The stars and arrow-heads in these figures are for illustrative purposes only; they indicate the start and current end points of the gesture being entered. Gesture entry starts when the user presses the mouse button and begins moving the mouse. When entering predictive gestures, users are simultaneously shown two forms of visual feedback. First, they see the gesture path they have entered so far, i.e. the gesture ink. Secondly, they see the gesture prediction drawn in red. The gesture

¹ For a video demonstration of the visual autocompletion styles please view the video accompanying this paper.

prediction is automatically scaled to match the shape and scale of the gesture ink. If a user chooses to stop entering the gesture at any time (by releasing the mouse button), then the currently predicted gesture is entered - as though the user drew the full gesture themselves.

Unlike the OctoPocus [5] and ShadowGuides [13] systems, SimpleFlow and SF-Path only show one gesture prediction at a time. If another prediction is more suitable, then the existing prediction is instantly switched out for the new prediction. Only one gesture prediction is shown because participants in our initial pilot studies voiced strong concerns about the high level of visual complexity that occurs when showing multiple simultaneous SimpleFlow predictions.

Another key reason for not showing multiple gesture predictions is we strongly suspect that the style of visual design used to show multiple gestures has a very significant impact upon user performance and preferences. The reason for this suspicion is due to research findings in psychology and perception around visual search, i.e. subjects search for a target stimulus (gesture) amongst distractor stimuli (predictions).

3.2 Standard Non-predictive Visual Feedback

The non-predictive style of visual feedback evaluated in this work is *Standard Feedback*. Standard Feedback is included because it is the typical gesture input mechanism and visual feedback used in many real-world gesture UIs, e.g. StrokeIt [12] for Windows, wayV [8] for Linux, FireGestures [14] for Firefox.

Standard Feedback does not give predictive visual feedback to users. When entering a gesture only the gesture ink is drawn on-screen. The gesture ink shows users the shape of the gesture they have entered, but no feedback about predictions are provided. Standard Feedback does not provide pre- or post-gesture information about whether the gestures are entered correctly or incorrectly.

For our experiment, during the course of Standard Feedback gesture input, gesture prediction does still take place but the predictions are not shown to the participants. This is done to ensure that the algorithm for recognising gestures is the same across all forms of predictive and non-predictive feedback. In this paper, Standard Feedback serves as a control for comparing and evaluating SimpleFlow and SF-Path.

3.3 SimpleFlow

SimpleFlow provides real-time visual feedback during gesture entry and always suggests a complete gesture. As soon as a user begins entering a gesture they start to receive continuously updated gesture predictions. A gesture prediction is drawn in red underneath the gesture ink (Figure 1). The predicted gesture is automatically scaled so it underlays the gesture ink and visually appears to continue the path of the gesture ink. Often the gesture ink and predicted gesture do not perfectly align on top of each other.

Figure 1 shows a full walk-through of SimpleFlow in action. In Figure 1(a), the user has started to enter their gesture, in this case, a triangle shaped gesture (Figure 4, Gesture 1). SimpleFlow suggests and displays a fully scaled triangle underneath the gesture ink. As the user continues to extend the edge of their input gesture (as shown in Figure 1(b)) the SimpleFlow feedback scales up to accommodate the change.

Figure 1(c) depicts the final stage of entering the triangle. The user draws the base of the triangle gesture, and decides that the full gesture prediction provided by SimpleFlow satisfies their needs. Now they cease their gesture input, safe in the knowledge that the correct complete gesture has been recognised.

SimpleFlow is scale and position invariant, however there is a limitation on the minimum size of the gesture predictions. If there was no minimum size limitation, the predicted gestures could be too small to see. For more details on the gesture prediction algorithm see the Appendix.

3.4 Scale Free Dynamic Paths

Scale Free Dynamic Paths (SF-Path) is like SimpleFlow. As with SimpleFlow, SF-Path is scale and position independent, and has an imposed limitation on the minimum size of the gesture predictions. The critical difference between SF-Path and SimpleFlow is how much of the predicted gesture is shown as visual feedback. As the results show, this visually subtle and small difference between SF-Path and SimpleFlow has a very significant impact on user performance and preferences.

Figure 2 shows the visual feedback provided while entering the triangle gesture with SF-Path. As with SimpleFlow the gesture ink is black, and the predicted gesture is shown in red. Unlike SimpleFlow, the complete gesture prediction is not shown on-screen by SF-Path. Instead the prediction and gesture ink are merged to form a single combined gesture, so that the gesture prediction looks to be a continuation of the gesture ink.

SF-Path is less visually complex than SimpleFlow, as there are no red lines underlying the gesture ink (Figure 2). When the gesture predictions are wrong the visual complexity of SimpleFlow is beneficial. For example, Figure 3 shows a side by side example of the differences between SimpleFlow and SF-Path when a visual prediction is wrong. In this example a user is trying to enter a triangle gesture. The prediction algorithm is mistakenly suggesting the gesture C. In Figure 3(a), with SimpleFlow, it is clear that the prediction algorithm is mistakenly suggesting the C gesture. Unfortunately, with SF-Path (Figure 3(b)) it is not at all clear which predicted gesture is merged with the gesture ink to form the combined gesture.

SF-Path resembles the visual feedback provided by OctoPocus [5] - if OctoPocus is altered to show one gesture suggestion at a time, its gesture suggestions are made scale and aspect invariant, and gesture abbreviation is allowed.

4 Experiment

Our experiment tests whether there are significant user performance and preference differences between predictive gestures and non-predictive gestures. It also establishes whether there are significant differences between the two styles of predictive visual autocompletion (SF-Path and SimpleFlow).

4.1 Hypothesis

For each of these hypotheses we are interested in understanding whether they do or do not hold true between the three styles of visual feedback and autocompletion, i.e.

Standard Feedback, SF-Path and SimpleFlow. For example, are SF-Path gestures faster than SimpleFlow gestures? We hypothesise that:

- **H1 Shorten:** Predictive gestures enable users to reduce gestures to a shorter form, by entering abbreviated gestures rather than full gestures. Like predictive text entry systems.
- **H2 Accuracy:** Predictive gestures with visual autocompletion improves the accuracy with which users enter gestures.
- **H3 Speed:** Inputting predictive gestures with visual autocompletion is slower.
- **H4 Cognitive Load:** Cognitive load is higher for predictive gestures. Visual autocompletion places higher cognitive demands on users during gesture entry.

For H1 Shorten we expect that the continuous predictive visual feedback provided during the course of gesture entry will enable users to significantly shorten the gestures they enter. If true, users will enter abbreviated short partial gestures rather than whole gestures. Of particular interest is how much do they shorten the gestures by, a small amount or large amount?

Accuracy is an important feature of gesture input. Providing effective real-time gesture feedback should improve users' ability to enter gestures correctly. Whether SimpleFlow and SF-Path are effective forms of visual feedback is established with the H2 hypothesis.

The speed with which gestures are entered is important. If gesture entry takes too long it could be distracting and interfere with workflows. With the H3 hypothesis we expect a speed / accuracy tradeoff, where speed is how long it takes to enter a gesture. Predictive gesture entry may be slower because users spend time evaluating and adapting their gesture input based on the real-time feedback.

Hypothesis 4 establishes whether predictive visual autocompletion requires more cognitive resources than no predictive visual feedback. Ideally, well-designed forms of visual feedback should not increase cognitive load during gesture entry.

4.2 Design

The experiment task was to input a series of gestures correctly. A gesture was deemed correctly entered when the gesture inputted by the participant matched the target gesture they had been instructed to enter. Whether a gesture matched was evaluated by the gesture recognition algorithm.

A within-subjects experiment design was used. The experiment was 3x1, where the independent variable was the style of visual gesture autocompletion and the dependent variable was gesture input accuracy. The three levels of the independent variable were Standard Feedback, SF-Path and SimpleFlow. While the two levels of the dependent variable were *Correct* or *Incorrect*, corresponding to whether the entered gesture correctly or incorrectly matched the target gesture.

All participants were presented with three randomly ordered blocks of randomly ordered gestures. Each block was presented once and corresponded to a level of the independent variable, i.e. Standard Feedback, SF-Path, or SimpleFlow. Within each block each gesture was repeated three times, distributed randomly within a block. At the start of each block participants practiced inputting three gestures. The visual autocompletion drawn during practice gestures was based on the level of the independent variable.

Participants completed the experiment in a single session. Each participant completed a short post-hoc questionnaire, where they provided their rank preferences for the visual autocompletion styles.

Participants. Eighteen volunteer participants took part, three of which did the pilot experiments and the remaining fifteen completed the main experiment. Of those fifteen 10 were male and 5 female. Participants' average age was 26.5 years, with a standard deviation of 3.8 years. All participants naturally used their right hand to control the mouse. The fifteen participants entered 2295 gestures, of which 135 were practice gestures.

Materials. Figure 4 shows the sixteen gestures used for the experiment. The gestures are from the work on the \$1 Recognizer [32]. We picked those gestures as they were independently created and have been used in other gesture experiments [32]. Using an independently created set of gestures helps avoid introducing an experiment bias, as the experiment did not run with a set of gestures specifically designed for the experiment task.

The gesture recognition algorithm was kept the same between the three levels. What differed between the levels was whether visual feedback was shown, and what kind of visual feedback it was. Details about the gesture recognition algorithm are provided in the Appendix.

Procedure. The experiment took place in a private room, free from external stimuli. A PC running Windows XP was used to run the experiment, and the same mouse was used by all participants. The PC monitor was equipped with a Tobii T60 eyetracker, which was used to capture pupil dilation at 60Hz. The eyetracker data was used to calculate cognitive load (Section 5.4). Pixels on the monitor were square.

Participants were first calibrated on the Tobii eye tracker, with Tobii's standard calibration and eye tracking software. Then each participant was given time to practice inputting gestures using the mouse, during which no predictions were displayed. The participants then practiced the stimulus-response gesture input task three times, with gestures from Figure 4. At the start of each block participants also practiced inputting gestures, during which they saw the visual autocompletion specific to the block.

The gesture input stimulus-response task consisted of presenting a target gesture to the participants, which was displayed centered on-screen for two seconds. During the two seconds the mouse pointer was not shown on-screen. Text above the target gesture instructed participants to look at the gesture. After the two seconds expired the screen was blanked, a text message appeared informing participants to draw the gesture, and the mouse pointer appeared centered on-screen. The mouse pointer was centered on-screen to prevent a carryover effect occurring between tasks, which could arise if the mouse pointer location was carried between gesture tasks.

Participants then inputted a gesture by pressing and holding the left mouse button down and moving the mouse. When the participant released the mouse button, the gesture was considered complete.

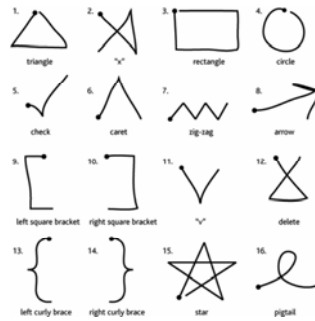


Fig. 4. Unistroke gestures used in the experiment, from [32]. Black points are the starting point for drawing the gestures.

Statistical Techniques. To analyse hypothesis H1 and H3 we use Balanced Repeated Measures Within-Subjects ANOVAs with Bonferroni pairwise comparisons. For H4 we first apply a standard log transform to the results, as tests of normality indicated the results had a non-normal distribution until transformed. Then we apply the same analysis techniques as for H1 and H3.

For hypothesis H2 we use Repeated Measures Logistic Regression (RMLR) with Bonferroni pairwise comparisons. RMLR is used because gesture accuracy is a categorical binary variable, i.e. gestures are either entered *Correctly* or *Incorrectly*.

A Bonferroni test compares each level with every other level, and establishes whether performance between the levels is significantly different. Balancing is performed by selecting a subset of results randomly distributed in the results. For example, if there are 600 results for Standard Feedback (Standard), 550 for SF-Path and 500 for SimpleFlow, then 500 results are randomly selected from each of the levels and pairwise comparisons are performed on those results.

Rather than reporting every p value, we report the meaningful p values, i.e. the highest significant p values out of each three way pairwise comparison. There are nine p values generated per set of results (Standard vs SF-Path, Standard vs SimpleFlow, SF-Path vs SimpleFlow) * (All, Correct, Incorrect gestures).

Accepting Gesture Predictions. For the blocks where participants saw predictive visual autocompletion (SF-Path and SimpleFlow), they were informed they could accept a gesture prediction without having to enter the full gesture. Accepting a gesture prediction was achieved by stopping gesture entry. Participants stopped gesture entry by releasing the mouse button. For example, in Figure 1 a participant could stop entering a gesture at time point (a) if they wanted to accept the prediction and input a triangle gesture.

5 Results

Overall, users of SimpleFlow automatically shorten gestures by 41%, while simultaneously improving the accuracy (+13%, from 68% to 81%) and speed (+10%) of

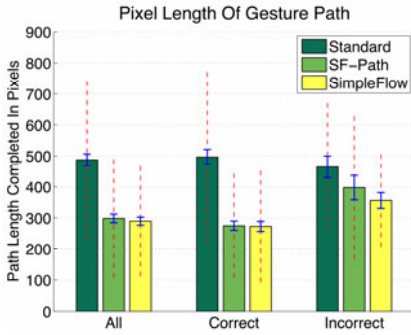


Fig. 5. Pixel path length of entered gesture. Error bars are 95% Wald confidence interval, dashed lines standard deviation

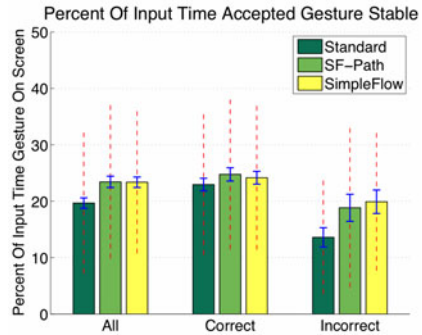


Fig. 6. How long was the accepted gesture stable on screen, as percentage of gesture input time

gesture input - and this is achieved with no significant increase in cognitive load. Results from the post-hoc questionnaire indicate that participants prefer SimpleFlow over SF-Path and Standard Feedback.

A total of 135 training gestures and 21 error gestures were removed from the results, leaving 2139 gestures available for analysis, with 714 gestures in Standard Feedback, 713 in SF-Path and 712 in SimpleFlow conditions. Gestures were classified as errors where participants accidentally clicked the mouse button and did not move the mouse.

5.1 Hypothesis: H1 Shorten

Users can and do take advantage of the gesture predictions. Participants enter 39%+ shorter gestures with SimpleFlow and SF-Path, hypothesis H1 is true. SimpleFlow and SF-Path successfully enable users to enter abbreviated short partial gestures rather than whole gestures. This implies that the predictive gesture algorithm makes accurate predictions early and continuously during the course of gesture entry.

Abbreviated Gestures. To establish whether participants abbreviate gestures, we analyse the pixel path length of entered gestures. Gesture path length is a measure of how many pixels are travelled when entering a gesture. Gesture path length is significantly ($p < 0.037$) shorter for All, Correct and Incorrect SF-Path (39%, All) and SimpleFlow (41%, All) gestures, compared to No Feedback (Figure 5). No significant difference exists between SF-Path and SimpleFlow. H1 is true, based on the gesture pixel path length.

A shorter gesture pixel path length does not definitively establish that participants abbreviate gestures. Shorter path lengths could mean participants enter smaller gestures. To rule out this possibility, we measure and analyse the entered path length divided by the predicted path length. We refer to this measure as the Partial Gesture Ratio (PGR). The lower the PGR the better. A $PGR < 1$ means the entered gesture is

Table 1. Path length of entered gestures, as PGR

	All	Correct	Incorrect
Standard	1.014 (0.288)	0.994 (0.057)	1.125 (0.535)
SF-Path	0.733 (0.395)	0.668 (0.314)	0.901 (0.542)
SimpleFlow	0.697 (0.323)	0.685 (0.297)	0.791 (0.393)

Table 2. Number of times gesture prediction changed

	All	Correct	Incorrect
Standard	6.51 (3.73)	6.01 (3.69)	7.74 (3.50)
SF-Path	5.84 (4.30)	5.28 (4.02)	7.82 (4.93)
SimpleFlow	5.48 (3.88)	5.13 (3.71)	6.94 (3.87)

shorter than the predicted gesture, $PGR > 1$ means the entered gesture is longer than the predicted gesture, and $PGR = 1$ means the gestures are the same length.

Based on the PGR we find that All, Correct and Incorrect gestures entered with SF-Path and SimpleFlow are significantly ($p < 0.002$) shorter than Standard Feedback (Table 1). This further confirms that participants do abbreviate gestures.

Early And Continuous Predictions. Do participants enter partial gestures because the prediction algorithm generates accurate predictions early and continuously during gesture entry? We analyse whether the gesture predictions stabilise early and stop changing. Early stabilisation is measured by dividing the time when gesture predictions stop changing by the time taken to input the gesture.

Predictions do stabilise early and stop changing during gesture entry (Figure 6). For All gestures Standard Feedback is stable for 39% of gesture input time, while SF-Path and SimpleFlow are stable for 47% of input time. All and Incorrect gestures are stable significantly ($p < 0.001$) longer for SF-Path and SimpleFlow. There are no significant differences for Correct gestures.

Number of Predictions. How often the predictions change before stabilisation could effect whether users enter partial gestures. Ideally a prediction does not change too often, as too many changes could make it harder for a user to judge which prediction is emerging as the winning prediction.

SF-Path and SimpleFlow generate 5.84 and 5.48 prediction changes for All gestures (5.28 and 5.13 for Correct), which are significantly less gesture prediction changes, for All ($p < 0.003$) and Correct ($p < 0.017$) gestures. No significant differences occur for Incorrect gestures. Shown in Table 2 are the mean number of times the gesture predictions changed.

5.2 Hypothesis: H2 Accuracy

By more than 10% both SimpleFlow and SF-Path significantly ($p < 0.003$) improve gesture entry accuracy, compared to Standard Feedback. Hypothesis H2 is true,

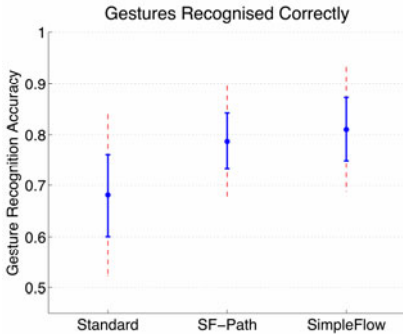


Fig. 7. Predictive gestures with visual auto-completion improves performance when entering gestures

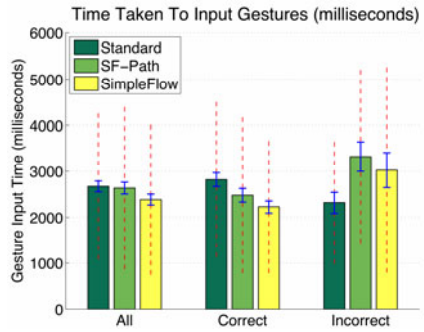


Fig. 8. Length of time taken to input gestures varies based on visual autocompletion

predictive gestures with visual autocompletion does improve the accuracy with which users enter gestures.

As shown in Figure 7, for Standard Feedback participants correctly entered gestures 68.1% (SD 15.8%) of the time. The gesture accuracy increased to 78.6% (SD 10.9%) and 81% (SD 12.2%) for SF-Path and SimpleFlow respectively. Performance is not significantly ($p < 0.05$) different between SF-Path and SimpleFlow. In Figure 7 the dot is the mean accuracy, the error bars are the 95% Wald confidence interval and the dashed lines the standard deviation.

5.3 Hypothesis: H3 Speed

Unexpectedly, SimpleFlow is more than quarter of a second faster (+10%, $p < 0.012$) for entering gestures than Standard Feedback and SF-Path. Entering gestures with SF-Path is as fast as with Standard Feedback (Figure 8). Hypothesis 3 is rejected, as visual autocompletion does not slow gesture input.

Time Taken To Enter Gestures. Focusing on the results for Correctly entered gestures - SF-Path is 344 milliseconds faster ($p < 0.006$) than Standard Feedback, and SimpleFlow is faster ($p < 0.038$) than both (Figure 8). H2 is rejected again.

When gestures are entered Incorrectly, then Standard Feedback is faster ($p < 0.009$) than both SF-Path and SimpleFlow. For Incorrect gestures H3 is true. No significant time differences exist between SF-Path and SimpleFlow for Incorrect gestures.

Rate Of Gesture Input. Is SimpleFlow the fastest input style because participants' rate of input is higher? When entering gestures how fast do users move the on-screen mouse pointer? Rate of input is calculated as the average number of pixels travelled by the mouse pointer per millisecond during gesture input.

Standard Feedback leads to a significantly ($p < 0.0001$) higher rate of gesture input than both SF-Path and SimpleFlow; for All, Correct and Incorrect gestures (Figure 9). There are no significant differences between SF-Path and SimpleFlow.

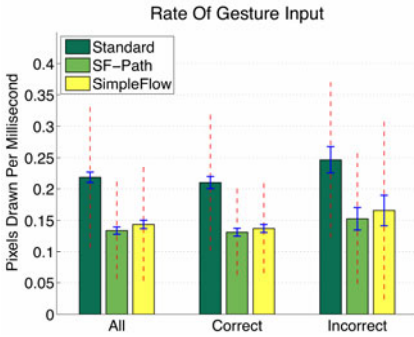


Fig. 9. Rate of input

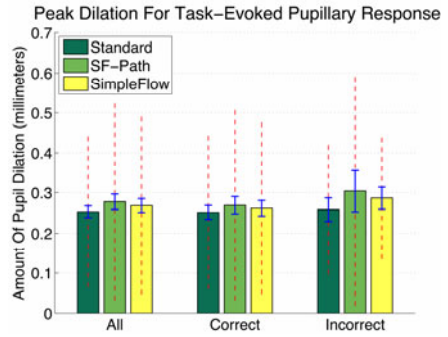


Fig. 10. TEPR Peak Dilation measures of Cognitive Load

This confirms that the speed improvement for SimpleFlow is not due to an improved rate of input. The speed improvement occurs because participants enter abbreviated gestures, which means it takes less time to enter the partial gestures.

5.4 Hypothesis: H4 Cognitive Load

Does the visual autocompletion place higher cognitive demands on users, than no visual feedback? Or does the visual feedback even decrease the cognitive load? Surprisingly H4 is rejected, SF-Path and SimpleFlow do not lead to increased cognitive load. There are no significant differences ($p < 0.05$) in cognitive load between Standard Feedback, SF-Path and SimpleFlow (Figure 10). This result applies to All, Correct and Incorrect gestures.

We measured cognitive load by measuring Task-Evoked Pupillary Response (TEPR), which is known to be a reliable indicator of cognitive load [7, 19, 21]. Task-Evoked Pupillary Response is a measure of how much the pupil dilates over time as a function of task hardness. The larger the change in pupil dilation the larger the cognitive load.

As recommended in the literature we removed blink artifacts, applied a low-pass filter to the dilation measures, used a 500ms baseline and used TEPR Peak Dilation to calculate cognitive load [7, 21]. The baseline was captured each time participants entered a gesture, right before they began entering the gesture. Measuring TEPR with a non-contact eyetracker enables us to measure real-time non-subjective quantifications of cognitive load, as an alternative to subjective work load assessments such as NASA TLX [17].

5.5 User Preferences

Participants preferred SimpleFlow 2.2 times more often than SF-Path, and 2.75 times more often than Standard Feedback (Figure 11). In the post-hoc questionnaire participants ranked Standard Feedback, SF-Path and SimpleFlow on a scale from 1 to 3, where 1 is most preferred and 3 least preferred. They could assign equal ranks.

Figure 11 shows the number times each rank was assigned to Standard Feedback, SF-Path and SimpleFlow. Participants ranked SimpleFlow 1st eleven times, SF-Path

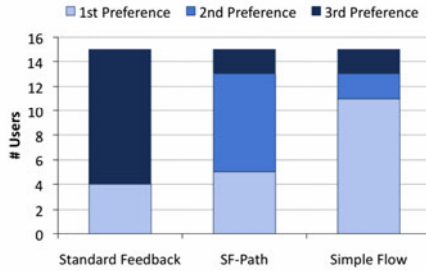


Fig. 11. User preferences as indicated in post-study questionnaire

1st five times, and Standard Feedback 1st four times. SF-Path is most often ranked 2nd, and Standard Feedback is most often ranked 3rd.

6 Discussion and Limitations

Our design aim for SimpleFlow, was to keep the complexity of the visual prediction feedback as simple as possible, while improving user performance. A related goal was to enhance gestural interaction without significantly changing it; in an effort to avoid requiring users to spend considerable amounts of time skilling up on a new interaction technique. Two further key aims were to enable users to take advantage of the gesture predictions without forcing them to use the predictions, and enable users to enter shorter gestures, rather than having to input full gestures. Looking at the results we find that SimpleFlow performs best out of all three forms of visual feedback, while also achieving our goals and aims. During the experiment users had few practice opportunities with the predictive gestures, yet they performed significantly better with predictive gestures versus non-predictive, while also preferring the predictions over no predictions.

Of particular interest we found that SimpleFlow is faster and more preferred to SF-Path, which is the same style of visual feedback as OctoPocus (when one gesture suggestion is shown at a time and the suggestions are scale and aspect invariant). This is interesting because the visual difference between SimpleFlow's visual feedback and SF-Path's visual feedback is very small, even though it has a critical impact. SimpleFlow shows a whole gesture suggestion in conjunction with the partial gesture a user has entered, while SF-Path completes the gesture for a user (Figure 3). When we followed up with users and informally asked them why they preferred SimpleFlow over SF-Path, the general consensus was that showing whole gestures clearly tells the users exactly which predicted gesture is getting matched with their partial gesture. While when partial gestures are completed in SF-Path, it is not always clear to the user why a gesture is completed the way it is, especially when the gesture prediction is wrong.

An interesting limitation of this work, which strongly suggests worthwhile future directions, is that we treat the users' performance and the performance of the gesture prediction algorithm as a combined system. This leaves open questions about the relationship between a user's performance, and what would happen if a better or worse gesture prediction algorithm is used? For example, if a perfect gesture recognition

algorithm is used would we find that users enter even shorter gestures? If so, what is the absolute lowest bound on how much users will abbreviate gestures by? Does that interact with the number of available gestures? Related questions include what, if any, are the effects of other performance characteristics of the gesture prediction algorithm, e.g. stability of predictions. Ultimately, what are the desirable characteristics of gesture prediction algorithms? Another interesting question arises around the set of gestures we ran the experiment with, as the range of gesture shapes could impact on user performance. For example, in Figure 4 gesture 2, 11 and 12 all share a common initial stroke (downward diagonal left stroke) - what would happen if a set of gestures is used that is optimised to minimize the shared starting strokes?

There are many other possibilities for future research, including extending SimpleFlow to handle multi-stroke gestures, creating UI techniques that enable users to interactively refuse and cancel a gesture suggestion, establishing what styles of pre- and post-gesture feedback is beneficial with visual predictions, and enabling users to quickly select from a set of gesture suggestions without having to complete the gestures. Finally, understanding more about why and how users decide to shorten predictive gestures would be very useful. Understanding this would help us further enhance and influence gesture abbreviation, i.e. enable users to enter even shorter gestures more quickly.

7 Conclusions

We found that users of predictive gestures with SimpleFlow and SF-Path visual feedback will significantly shorten gestures (like predictive text entry systems), with no significant increase in cognitive load. Further, SimpleFlow successfully enhances users' gesture entry, both speeding it up and improving the accuracy; along with users preferring SimpleFlow the most. The visual feedback provided by SimpleFlow is visually simple and minimal, and could easily be added to existing gesture entry systems without requiring significant changes to them.

References

1. Accot, J., Zhai, S.: Beyond Fitts' Law: Models for trajectory-based HCI tasks. In: Proc. CHI 1997, pp. 295–302 (1997)
2. Agar, P., Novins, K.: Polygon recognition in sketch-based interfaces with immediate and continuous feedback. In: Proc. GRAPHITE 2003, pp. 147–150 (2003)
3. Appert, C., Zhai, S.: Using strokes as command shortcuts: Cognitive benefits and toolkit support. In: Proc. CHI 2009, pp. 2289–2298 (2009)
4. Arvo, J., Novins, K.: Fluid sketches: Continuous recognition and morphing of simple hand-drawn shapes. In: Proc. UIST 2000, pp. 73–80 (2000)
5. Bau, O., Mackay, W.E.: Octopocus: A dynamic guide for learning gesture-based command sets. In: Proc. UIST 2008, pp. 37–46 (2008)
6. Baudisch, P., Cutrell, E., Robbins, D., Czerwinski, M., Tandler, P., Bederson, B., Zierlinger, A.: Drag-and-Pop and Drag-and-Pick: Techniques for accessing remote screen content on touch- and pen-operated systems. In: Proc. Interact 2003, pp. 57–64 (2003)

7. Beatty, J., Lucero-Wagoner, B.: The Pupillary System. In: *Handbook of Psychophysiology*, 2nd edn., ch. 6, pp. 142–161. Cambridge University Press, Cambridge (2000)
8. Bennett, M.: wayv: Gestures for Linux (2011), <http://www.stressbunny.com/wayv>
9. Cao, X., Balakrishnan, R.: Evaluation of an on-line adaptive gesture interface with command prediction. In: *Proc. GI 2005*, pp. 187–194 (2005)
10. Cao, X., Zhai, S.: Modeling human performance of pen stroke gestures. In: *Proc. CHI 2007*, pp. 1495–1504 (2007)
11. Collomb, M., Hascoet, M., Baudisch, P., Lee, B.: Improving drag-and-drop on wall-size displays. In: *Proc GI 2005*, pp. 25–32 (2005)
12. Doozan, J.: Strokeit gestures for Windows (2011), <http://www.tcbmi.com/strokeit>
13. Freeman, D., Benko, H., Morris, M.R., Wigdor, D.: Shadowguides: Visualizations for in-situ learning of multi-touch and whole-hand gestures. In: *Proc. Tabletop 2009*, pp. 183–190 (2009)
14. Gomita.: Firegestures: Gesture control for Firefox web browser (2011), <http://www.xuldev.org/firegestures>
15. Gustafson, S., Bierwirth, D., Baudisch, P.: Imaginary interfaces: Spatial interaction with empty hands and without visual feedback. In: *Proc. UIST 2010* (2010)
16. Harrison, C., Tan, D., Morris, D.: Skinput: Appropriating the body as an input surface. In: *Proc. CHI 2010*, pp. 453–462 (2010)
17. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: *Human Mental Workload*, pp. 239–250. North Holland Press, Amsterdam (1988)
18. Igarashi, T., Matsuoka, S., Kawachiya, S., Tanaka, H.: Interactive beautification: A technique for rapid geometric design. In: *Proc. UIST 1997*, pp. 105–114 (1997)
19. Iqbal, S.T., Zheng, X.S., Bailey, B.P.: Task-evoked pupillary response to mental workload in Human-Computer Interaction. In: *Proc. CHI 2004, Extended Abstracts*, pp. 1477–1480 (2004)
20. Isokoski, P.: Model for unistroke writing time. In: *CHI 2001*, pp. 357–364 (2001)
21. Klingner, J., Kumar, R., Hanrahan, P.: Measuring the task-evoked pupillary response with a remote eye tracker. In: *Proc. ETRA 2008*, pp. 69–72 (2008)
22. Kristensson, P.O., Zhai, S.: Command strokes with and without preview: Using pen gestures on keyboard for command selection. In: *CHI 2007*, pp. 1137–1146 (2007)
23. Kurtenbach, G., William, B.: The limits of expert performance using hierarchic marking menus. In: *Proc. CHI 1993*, pp. 482–487 (1993)
24. Li, J., Zhang, X., Ao, X., Dai, G.: Sketch recognition with continuous feedback based on incremental intention extraction. In: *Proc. IUI*, pp. 145–150 (2005)
25. Li, Y.: Protractor: A fast and accurate gesture recognizer. In: *Proc. CHI 2010*, pp. 2169–2172 (2010)
26. Long Jr., A.C., Landay, J.A., Rowe, L.A., Michiels, J.: Visual similarity of pen gestures. In: *Proc. CHI 2000*, pp. 360–367 (2000)
27. MacKenzie, I.S., Soukoreff, R.W.: Text entry for mobile computing: Models and methods, theory and practice. *Journal of Human-Computer Interaction* 17, 147–198 (2002)
28. Rubine, D.: Specifying gestures by example. *SIGGRAPH Computer Graphics* 25(4), 329–337 (1991)
29. Soukoreff, R.W., MacKenzie, I.S.: Metrics for text entry research: An evaluation of msd and kspc, and a new unified error metric. In: *Proc. CHI 2003*, pp. 113–120 (2003)

30. Tandler, P., Prante, T.: Using incremental gesture recognition to provide immediate feedback while drawing pen gestures. In: Proc. UIST 2001, pp. 18–25 (2001)
31. Ward, D.J., Blackwell, A.F., MacKay, D.J.: Dasher - a gesture-driven data entry interface for mobile computing. In: Proc. UIST 2000, pp. 129–138 (2000)
32. Wobbrock, J.O., Wilson, A.D., Li, Y.: Gestures without libraries, toolkits or training: A \$1 recognizer for user interface prototypes. In: Proc. UIST 2007, pp. 159–168 (2007)
33. Yatani, K., Partridge, K., Bern, M., Newman, M.W.: Escape: A target selection technique using visually-cued gestures. In: Proc. CHI 2008, pp. 285–294 (2008)

Appendix: Gesture Prediction Algorithm

For experimental reproducibility this appendix outlines the gesture prediction algorithm (Figure 12), which is composed of a gesture concatenation algorithm and a gesture recognition algorithm. When a user enters part of a gesture, the gesture concatenation algorithm generates a complete gesture based on the partially entered gesture. Then the complete gesture is sent to a gesture recognition algorithm, to check whether it matches any of the trained gestures. The best matching gesture is used as the gesture prediction.

Gesture Concatenation Algorithm. The gesture concatenation algorithm is easy to implement, though the computational efficiency of it is open to improvement. The algorithm assumes that gestures start from the same points, like the \$1 Recognizer [32]. Unlike OctoPocus [5] and ShadowGuides [13] however, the algorithm is scale independent. Like \$1 Recognizer and OctoPocus, and unlike ShadowGuides, our algorithm is for single stroke continuous gestures. The algorithm also handles different gesture aspect ratios, like Protractor [25], but like OctoPocus and ShadowGuides it is not rotational invariant to gesture orientations. Below are the seven steps in the algorithm (see Figure 12):

1. During gesture entry, continuously capture the path of the incoming gesture, generating a partial gesture P (Figure 12(a)).
2. Measure the width w , height h and calculate the length l of the partial gesture P (Figure 12(b)).
3. Scale the trained gestures (Figure 12(c)) so they share the same width w and height h as the partial gesture P. We refer to this set of scaled gestures as SC (Figure 12(d)).
4. For each scaled gesture in SC (Figure 12(d)), remove a path length l from the start of each gesture. This generates a set of cropped gestures CG (Figure 12(e)).
5. Merge the partial gesture P with every cropped gesture in CG. This generates a new set of gestures PG, as shown in Figure 12(f). Each gesture in PG is a gesture prediction. The merging should be done such that the end of P is merged with the start of each cropped gesture in CG.
6. If the gesture recognition algorithm is scale dependent, then the gestures in PG may need to be rescaled to match the scale of the gestures used to train the gesture recognition algorithm.
7. To find the winning gesture prediction send each gesture prediction in PG to the gesture recognition algorithm. The best scoring gesture in PG is the gesture prediction.

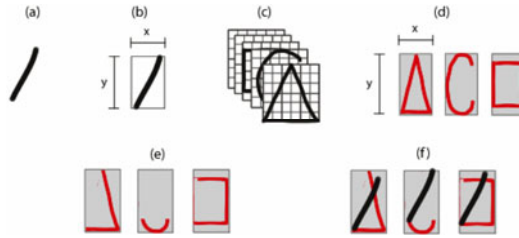


Fig. 12. Gesture Concatenation: (a) Current state of the user gesture is sampled; (b) height, width & length of the sample is measured; (c) the training gesture templates; (d) scale the training templates to match the partial gesture; (e) remove the sampled gesture from each of the scaled templates; (f) a is merged with each of e to produce new templates.

We imposed a constraint on the above algorithm, such that the predicted gesture cannot be scaled below a specific size. Without that constraint gestures could be scaled to only a few pixels wide or high, which would make the gestures visually indistinguishable.

Gesture Recognition Algorithm. The gesture recognition algorithm used is a scale and aspect invariant template matching algorithm. The code for which is based on wayV's gesture recognition algorithm [8]. wayV's algorithm accounts for bounding box issues that can arise where the input starts off purely vertical or horizontal.

The Perception of Sound and Its Influence in the Classroom

Sofia Reis and Nuno Correia

CITI and DI, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
2829-516 Caparica - Portugal
se.reis@fct.unl.pt, nmc@di.fct.unl.pt

Abstract. In this paper we describe a game to assess if the quantitative and graphical perception of sound by students can influence how they behave in the classroom. The game captures sound and shows the sound wave or the frequency spectrum, integrated with an animated character, to students in real time. The quieter the students are the higher the score. A survey was conducted to teachers from an elementary and secondary school to determine if they considered that noise, caused by the students, was a problem. Most of the teachers considered that students make too much noise. All the classes where the game was tested became quieter, thus showing that when these students perceived, in a quantitative way, how much their behavior was disruptive they were more inclined to be quiet or, at least, to reduce the amount of noise.

Keywords: game, persuasive technology, noise, classroom.

1 Introduction

Students can develop several different types of work in the classroom. Sometimes the teacher may expose something to the students and need the whole class to keep quiet and listen to what is being explained. Other times, students engage in tasks where they have to talk to each other. Nevertheless, whatever the activity is, if the sound level is too high it will be difficult for people to listen to each other. Furthermore, excessive noise can disturb students in other classrooms. Even though classroom noise may have different sources, some internal and others external to the classroom, here the focus is on the noise caused by the students talking to each other.

Noisy classrooms have an adverse effect on students' learning and the strain on teachers' voices can result in illness, as is described in Section 2.

We developed a game that motivates students to make less noise in the classroom. Games have already been previously used to change people's behavior (Section 3). Our game, which is further described in Section 4, is populated by characters that are only happy in the silence. If students are quiet they will gain more points. The game is, therefore, an incentive for students to change their behavior via an increased awareness of how much noise they are causing. The game was tested in an elementary and secondary school. Before the game was tested we assessed the school's initial situation to determine the teachers' and students' views about noise in the classroom

(Section 5). The game's test results are presented in Section 6. Finally, the conclusions and future work are in Section 7.

2 Effects of Classroom Noise on Students and Teachers

Classroom noise is detrimental to students' learning and can result in several adverse consequences. Noise decreases word recognition performance [1]. According to another source noise negatively affects performance in verbal tasks, like reading and spelling, and performance on speed tasks [2]. Noise is also prone to cause fatigue and headaches [3]. Noise is still related to the annoyance of both students and teachers. Chatter in the classroom is considered an annoying sound source. Teachers are more sensitive to noise than students and experience a higher level of stress. Females felt that noise caused them more stress than males [4].

It, therefore, seems to be beneficial if the noise level is low. To control the noise produced by students, inside the classroom, the teacher can resort to several classroom management strategies. The teacher can establish rules so that students know when they can talk and when they are supposed to be quiet and then reward students for adequate behavior by giving them a better grade, stars, points or smiley faces or showing their names in an honor board. A student's bad behavior can be punished by asking that student to leave the classroom or by giving her or him extra tasks to do or by keeping the student inside the classroom during recess [5]. However, none of this clearly shows to the students how much their behavior is disruptive. Perhaps, if students could perceive, in a quantitative way, how much their behavior is disruptive, then perhaps they would be more inclined to be quiet or to lower their voices. Our game was designed to test this hypothesis, as is further explained in Section 4.

Students are not the only ones negatively affected by a noisy classroom. Teachers, in consequence of the strain to their voices, may suffer health problems. Voice is one of the most important tools for a teacher because they have to talk for a long length of time and may also have to make themselves heard over a loud background noise. 62,7% of teachers are affected by voice problems [6]. Voice problems are more frequent in teachers than in other professions [7].

Voice problems can significantly affect a teacher. Teachers with frequent voice problems have decreased control and influence at work, low social support, poor job compensations, poor health and vitality perceptions and deficient job satisfaction [6]. Voice problems are also the cause behind lost days of work due to sick leave [8].

Voice problems are associated with the personality of the individual [9]. Teachers with voice problems tend to have a higher reactivity to stress. Still, if reactivity to stress is indeed a cause of voice problems, this personal characteristic may be difficult to change.

To ease or prevent voice problems the teacher can use a microphone for voice amplification. Adequate voice training, better acoustic conditions in the classroom and the absence of environmental irritants like dust or smoke are helpful. A fewer number of lessons will reduce the strain to the voice. Another solution is to reduce the number of children in the classroom as fewer children will produce less noise.

Usually, the teacher cannot change the work schedule and also cannot decide how many children are in the classroom. However if all children are quiet, while the teacher is explaining something, the teacher will not have to talk so loud, thus reducing the strain on the vocal cords. Our game motivates children to keep quiet or lower their voices. In the next section some examples of persuasive technology are presented.

3 Changing Behavior with Games

Persuasive technology has already been successfully used to change people's behavior. Here are a few diverse examples: help children deal with bullying situations [10]; motivate people to recycle waste materials [11, 12]; encourage children to decrease energy consumption at home [13]; encourage healthy dietary behaviors in kindergarten children [14]; raise teenagers' oral health and dental hygiene awareness [15]; stimulate physical play [16]; help people quit smoking [17]; help elders take their medication on time [18]; improve engagement in science controversies and develop skills in evaluating evidence and forming arguments [19]; improve awareness of drugs abuse effects [20]; raise awareness about water scarcity [21]; improve workers' mental engagement in routine activities [22]; treat cockroach phobia [23]; incentive people to throw rubbish in a bin, instead of on the floor [24]; incentive people to obey the speed limit [25].

These examples show that there is great potential to alter people's behavior, not only in the classroom, but in many other situations.

In what particularly refers to influencing the amount of sound produced by the students while talking to each other, in [26] a sound level meter was used to monitor a free study period. An observer recorded the data from a position in the rear center of the room and wrote it on sheets of paper attached to a clipboard. If the students kept the sound level low, for ten minutes, they would receive two extra minutes for the gym period and a two minutes break to do whatever they wanted, before the beginning of the next ten minutes period. However, if the students became too noisy, during the ten minutes silent period, a whistle would be blown and the timer would be reset back. In [27], the authors resorted to an automated clown to show the children if the sound level was too high. The clown had five lights that simulated the jacket buttons, two lights that simulated the eyes, one light that simulated the nose and five lights that simulated the mouth. If the children kept the sound level low the clown's lights would turn on. James W. Groff patented an alarm that emits a sound when the classroom is too noisy [28].

4 The Game

Our game is a game that shows students, in real time, quantitatively and graphically, the amount of sound in the classroom. The game runs in a computer that is connected to a video projector or to an interactive whiteboard so that all students can see the output of the game. Sound is captured through a microphone connected to the computer where the game is running. The game is populated by characters that enjoy

the silence. If the amount of sound the microphone is detecting is low the score increases. If the amount of sound the microphone is detecting is too high the score decreases and may even become negative.

The game's name is "The Castle of Count Pat". The characters of the game are Pat, a centuries old vampire, Pat's Cat and the Moon. The game begins with a story that explains the necessity of peace and quiet at Count Pat's Castle (Fig. 1). Count Pat does not like noise because he wants to sleep peacefully in his coffin. Count Pat's Cat likes silence because noise scares away the rats. It is difficult for Pat to adapt to nowadays noisy times so he took a decision: everyone who is making noise shall be turned into a vampire. If students want to keep their lives, they will have to keep quiet or at least lower their voices. The story is presented as a sequence of images. The teacher reads the story aloud to the students. To advance, the teacher presses the next button at the bottom of each image of the story. We chose this sort of interface because, this way, the teacher can present the story to the students at her or his desired speed. A video would make it more difficult to slow the pace of the story if some of the students had not understood it.

After the story in Fig. 1, the game starts. The teacher, together with the students, can choose a character to interact with: Count Pat (Fig. 2 (a)), Count Pat's Cat (Fig. 2 (b)) or the Moon (Fig. 2 (c)).

The sound wave is below Count Pat and the Moon. The frequency spectrum, after a Fourier transform is performed on the sound data, is integrated in the Cat's fur. At the bottom left corner of the game's interface the three available characters are presented. The teacher, or one of the students, can change the character by clicking on one of them. The other elements of the interface are: a volume bar that shows the amount of sound the microphone is detecting; the score; a button to end the game; and a button to pause the game.

Each character has three states. In state 1 all the characters are sleeping (Fig. 2). This is the best possible state. While in state 1 the score increases 0.2 points a second. The character remains in state 1 while the volume bar shows only its green section. The volume bar is divided in three sections. The leftmost section is green. The middle section is yellow. The rightmost section is red. If the volume bar shows the green and yellow sections, the character changes to state 2. Here, the score increases 0.1 points a second. In state 2 Pat opens the coffin, the Cat stands up and the Moon awakes (Fig. 2). If the volume bar is showing the green, yellow and red sections the character changes to state 3. This means the students are making too much noise. While the character is in state 3 the score decreases 0.1 points a second. If the students do not become quieter the score can turn negative. In the state 3, the characters are wide awake and angry. Pat puts his hands over his ears, the Cat's fur stands on end and the Moon shows a displeased face (Fig. 2).

Previous tests were conducted, in a classroom, with the students and their teachers, to decide when the amount of sound detected by the microphone is too high in order to determine the green, yellow and red sections of the volume bar. We were told, by the teacher, when the noise was considered excessive.










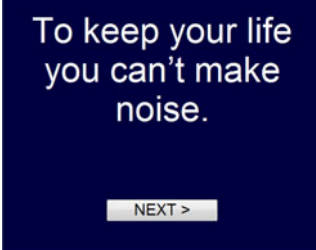
- (1) 
- (2) 
- (3) 
- (4) 
- (5) 
- (6) 
- (7) 
- (8) 
- (9) 
- (10) 

Fig. 1. The game begins with a story that explains why silence is necessary at Count Pat's Castle

To determine the state of the character we calculated the average of the last 10 activity level values of the microphone. The activity level is the amount of sound the microphone is detecting. Values range from 0, where no sound is detected, to 100, where very loud sound is detected. Each second the game collects 24 values of the microphone's activity level. We calculated the average of the last 10 activity level values because, if individual values were considered, the character would change between states too fast. At each second, the last computed average value is evaluated to determine how much to increase the score.

While the game is paused the score does not change. After the game ends the final score is displayed.

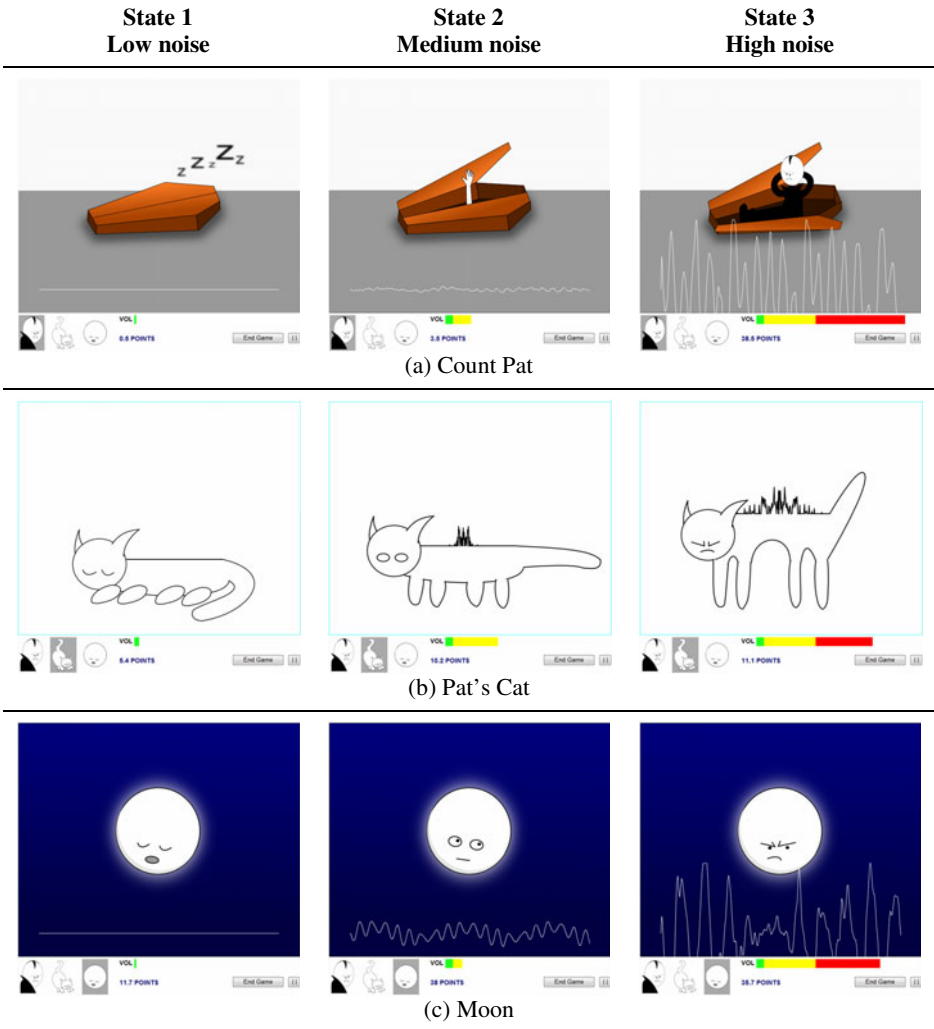


Fig. 2. The different states of the three characters

5 Assessment of the School Where the Game Was Tested

The game's tests were conducted in a school that is both an elementary and a secondary school. To evaluate the school's initial situation we conducted a survey to 60 of the school's 150 teachers. The teachers were questioned, with a paper survey, at the staff room, a place where teachers can rest. The survey was previously tested and was anonymous, but we were present at the staff room in case teachers had any doubt about the survey.

Through the survey we tried to determine if noise, caused by the students, was a problem. 78,3% of the inquired teachers agreed or strongly agreed that students make too much noise and that negatively influences their learning (Fig. 3).

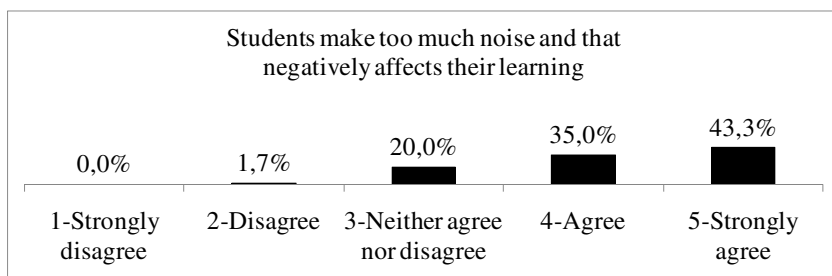


Fig. 3. Do teachers think that students make too much noise?

Almost all teachers agreed, or strongly agreed that when the students are making too much noise it is more difficult for them to teach the class (Fig. 4).

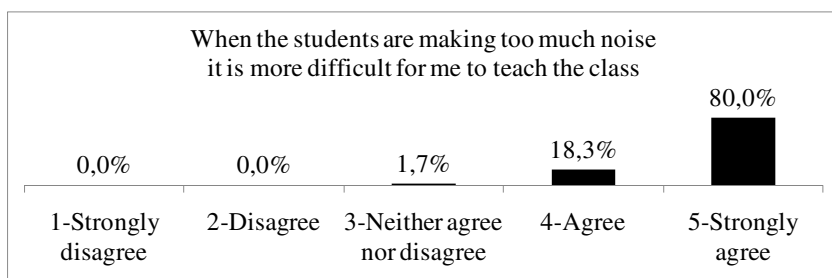


Fig. 4. Do teachers find it more difficult to teach the class when the students are making too much noise?

45% of the teachers reported voice problems like a hoarse voice, pains, vocal cord nodules or polyps or even being completely aphonic. Two of the enquired teachers underwent surgery because of their voice problems. One teacher afflicted by voice problems mentioned that those voice problems resulted in a depression and another reported feeling extremely tired.

Even though almost half of the enquired teachers have voice problems, 45% is not a percentage as high as the one found in [6].

We asked teachers if they thought that students were aware that talking while the teacher is explaining something negatively affects their learning (Fig. 5). 25% of the enquired teachers strongly disagree or simply disagree that students are aware of it. 35% neither agree nor disagree. 40% agree, or strongly agree, that students are conscious that their behavior is negatively affecting their learning.

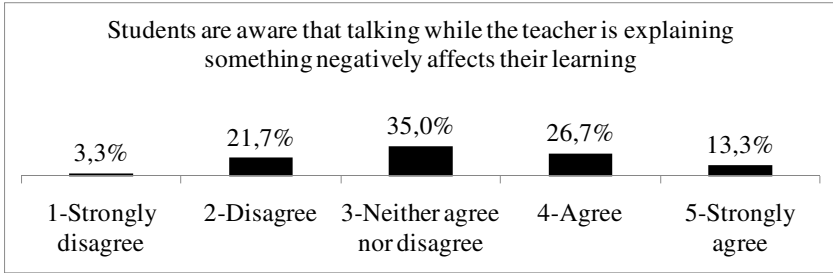


Fig. 5. Do teachers think that students are aware that talking while the teacher is explaining something negatively affects their learning?

When we asked the teachers if they thought that showing the students how much noise they are making would cause them to be quieter, 48,3% of the teachers agreed or strongly agreed this would work (Fig. 6). 36,7% of the teachers think that showing the students how much noise they are making would have no effect and 15% think that students would make even more noise. So, even though 40% of the enquired teachers believe that students are aware that they make too much noise in the classroom, 48,3% of the teachers believe that showing the students how much noise they are making is an extra reinforcement that might have some positive effect.

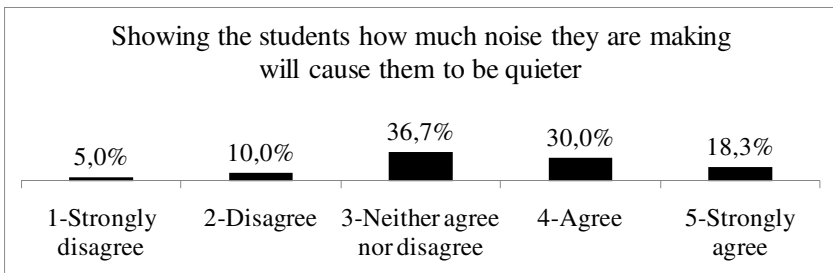


Fig. 6. Do teachers think that showing the students how much noise they are making will cause them to be quieter?

We tried to determine if the teachers’ views about noise in the classroom were similar to the students’ views. To this effect we enquired 81 students from 4 classes. These classes are the same classes where the game was tested. The survey was anonymous and previously tested. Only 34,5% of the inquired students think that their

colleagues make too much noise in the classroom (Fig. 7). In contrast, 78,3% of the inquired teachers consider the noise caused by the students excessive. So, the inquired students and teachers have different views and perhaps students do not completely realize that, according to the teachers, they are too noisy.

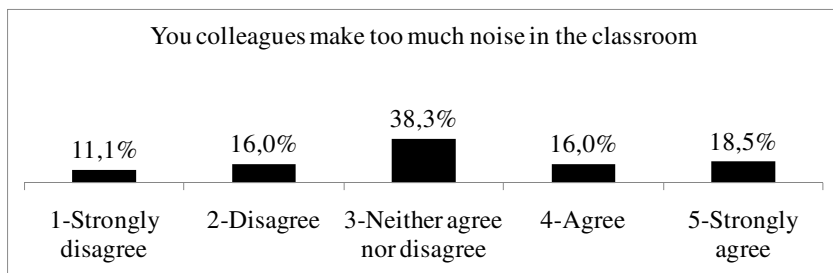


Fig. 7. Do students think their colleagues make too much noise in the classroom?

However, 61,1% of the students agree or strongly agree that if students are quiet, the grades will be better (Fig. 8). Therefore, most students and teachers agree that a quieter class will result in better grades.

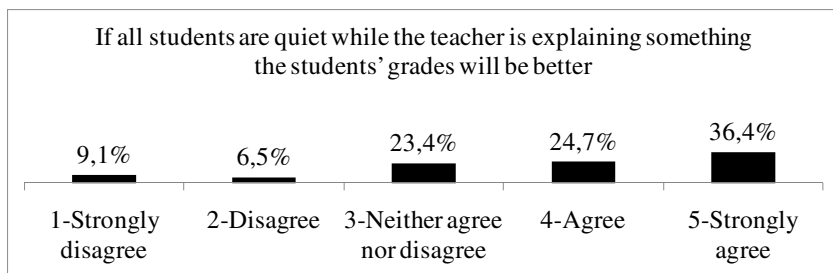


Fig. 8. Do students think that if they are quiet the grades will improve?

6 Testing the Game

Our game was tested in four classes of an elementary and secondary school. We shall refer to the classes as classes A, B, C and D. All of them were practical classes of Informatics. The classes' duration was 90 minutes. The number of students, grade, age average and age range of each class can be observed it Table 1.

Table 1. Composition of the classes where the game was tested

Class	Teacher	Number of students	Grade	Age average	Age range
A	X	18	8	14.6	12-16
B	Y	27	9	14	13-17
C	X	21	9	15.5	15-17
D	X	15	12	17	16-18

Before any tests were conducted we interviewed the teachers of the classes to determine what they thought about them. Among teacher X’s classes, class D was expected to be the quieter one. Class D’s students were described by the teacher as being mature and hard working. Class A and class C’s students were younger and the teacher expected them to be noisier. Class A was considered problematic because, even though it had only 18 students, many of them were repeating that grade. The teacher told us she left that class often feeling tired and with a hoarse voice.

Teacher Y expected class B to be noisy because this was the class with the greatest number of students and also because the students were still very immature.

Both teachers told us that students often loose conscience of how loud they are talking when they are engaged in group work.

First, we computed the average amount of sound detected by the microphone in all the classes without the game (Fig. 9).

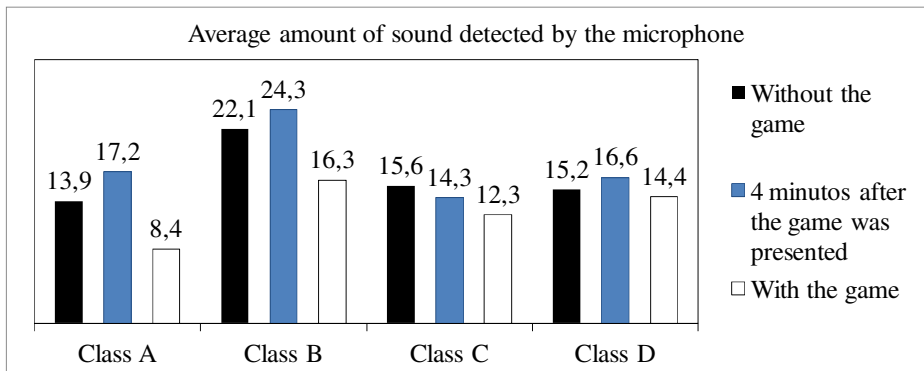


Fig. 9. Average amount of sound detected by the microphone with and without the game

Teacher X told us she was surprised that class A was the least noisy of her classes. The teacher attributed the greater fatigue, felt in class A, to the extra effort she had to make to motivate the students.

In the second lesson the game was tested. A video projector was used to show the game to all students (Fig. 10). The classroom had two whiteboards. We projected the game on the rightmost whiteboard. The teacher used the leftmost whiteboard when necessary. During the test of the game none of the teachers needed to show slides to the students. If this was the case, then a second video projector would be necessary.

At the beginning of the class the game was explained to the students. The students were told that the game would be tested in several classes and that the quieter class would receive a mystery gift. The three characters were shown to the students and they chose one of them. Afterwards, the lesson continued as usual.



Fig. 10. Class while the game is being tested

In all of the classes the amount of sound detected by the microphone decreased (Table 2). In class A there was a 39,6% decrease. This was the highest reduction of all the classes.

Table 2. Average amount of sound detected by the microphone during a first lesson without the game and during a second lesson with the game

Class	Average amount of sound detected by the microphone		Decrease
	Without the game	With the game	
A	13,9	8,4	39,6%
B	22,1	16,3	26,2%
C	15,6	12,3	21,2%
D	15,2	14,4	5,3%

Class D was the class where the use of the game resulted in the smallest decrease. The average amount of sound detected by the microphone decreased only 5,3%. In classes B and C there was a decrease of 26,2% and 21,2% respectively.

Also, in all of the classes, except in class C, the average amount of sound detected by the microphone, during the 4 minutes after the game was explained to the students, was higher than in the first lesson without the game (Fig. 9). This happened because, at first, the game captivated the students' curiosity and they tried to test it. Some students would raise their voices or whistle to see how the game reacted. They asked how much the score would increase, when they were quiet, and commented on the changes in the sound wave or in the frequency spectrum according to the different types of sounds produced.

Gradually, students turned their attention to their tasks. However the game was not forgotten. Throughout the class, students would often look at the projection to check the score. If one of the students was talking too loud, another one would usually ask her or him to lower her or his voice. Some students asked us what the score in the other classes was. For these students, competition with the other classes seems to be a motivation. Other students tried to set goals. Those students would turn to the rest of the class and say that they had to increase a certain number of points till the end of the

lesson and urged the others to be quiet. This type of peer pressure was more frequent in classes A and B than in classes C and D.

At the end of the lesson, where the game was tested, the students filled an anonymous survey that was previously tested.

In the survey, we asked the students if the class was quieter, during the use of the game (Fig. 11). In classes A, B and C most of the students agreed or strongly agreed that the classroom was quieter during the use of the game. These results are consistent with the results presented in Fig. 9. Class A is the class where more students agreed or strongly agreed that the class was quieter during the use of the game. The average amount of sound detected by the microphone decreased 39,6% in this class and this was perhaps easily noticed by most of the students. In class D, only 42,9% of the students agreed or strongly agreed that the class was quieter during the use of the game. This happened, probably, because the average amount of sound detected by the microphone decreased only 5,3% and this was hardly noticed by the students.

During the use of the game the class was quieter

1 - Strongly disagree; 2 - Disagree; 3 - Neither agree or disagree; 4 - Agree; 5 - Strongly Agree

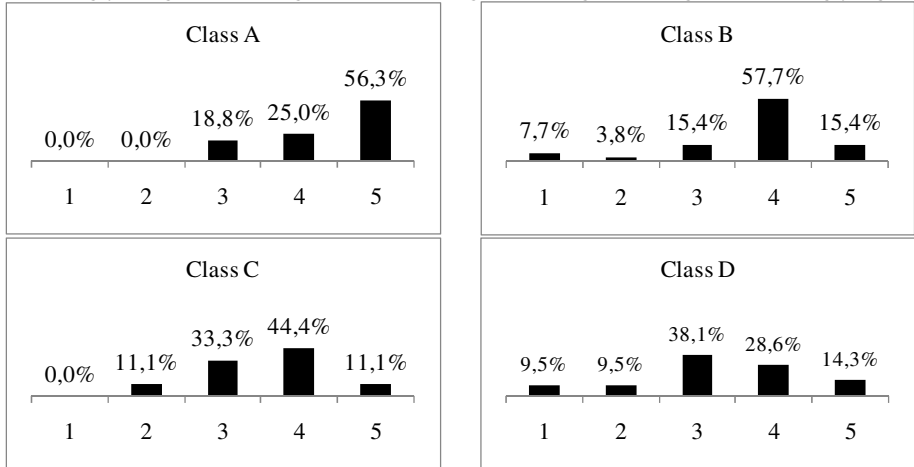


Fig. 11. Was the class quieter during the use of the game?

We wondered if the game would cause the students to pay more attention to the class. The game could also have the opposite effect. If the students spent a lot of time looking at the game that could reduce the time they spent listening to the teacher or working in their tasks. So, we asked the students if the game helped them pay more attention to the lesson (Fig. 12). Only in class A did most of the students agreed or strongly agreed that the game helped them pay more attention to the class. In classes B, C and D the answer “Neither agree or disagree” was the most chosen one. We interviewed the teachers of classes A, B, C and D and they noticed no visible change in the students’ attention to the lesson, even in class A. This seems to indicate that the game did not significantly influence the attention of the students to the lesson.

The game helped me pay more attention to the lesson

1 - Strongly disagree; 2 - Disagree; 3 - Neither agree or disagree; 4 - Agree; 5 - Strongly Agree

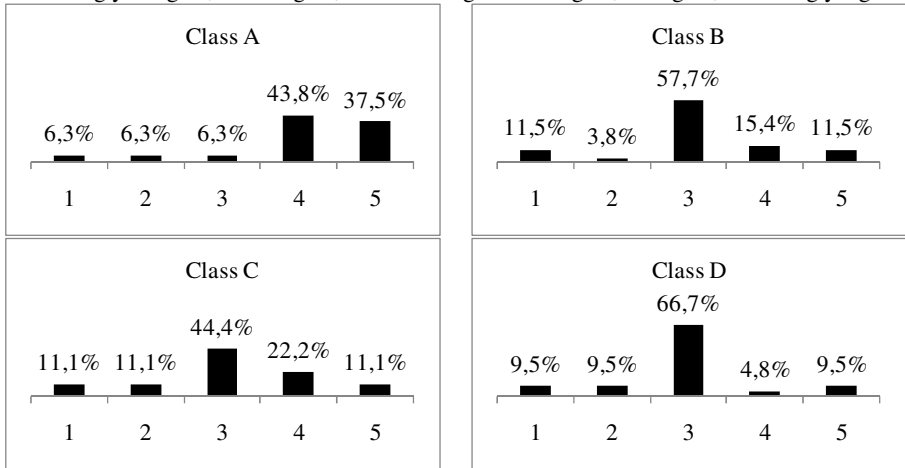


Fig. 12. Did the game helped students pay more attention to the lesson?

To make clear if the game decreased the students' attention to the teacher we included another question in the survey. We asked the students if the game distracted them from what the teacher was explaining (Fig. 13).

The game distracted me from what teacher was explaining

1 - Strongly disagree; 2 - Disagree; 3 - Neither agree or disagree; 4 - Agree; 5 - Strongly Agree

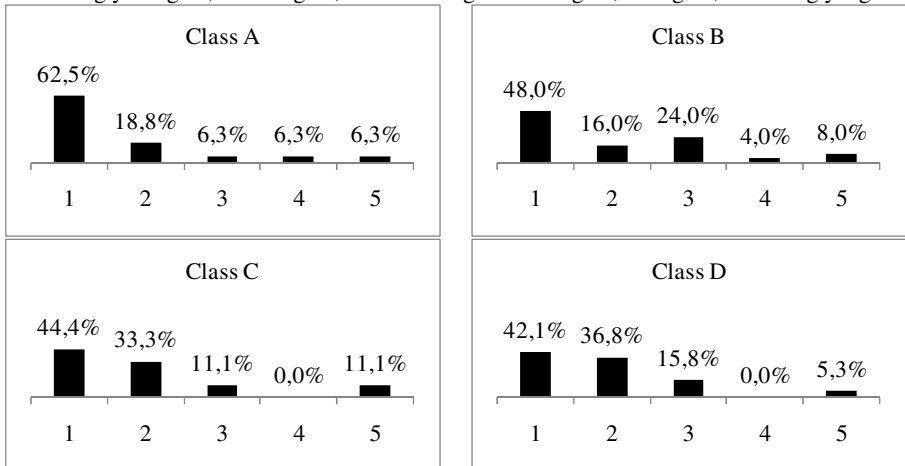


Fig. 13. Did the game distract the students from what the teacher was explaining?

In all the classes, the attention to what the teacher was explaining of most students was not negatively affected by the game. The percentage of students that agreed or strongly agreed that the game distracted them is quite small in all the classes. This indicates that even though the game did not cause the students to pay more attention it also did not cause the opposite effect. As the game does not measure the attention of the students to the lesson we consider that these results are not surprising.

In the survey we asked the students if it was important for them to obtain a good score in the game (Fig. 14). In classes A and B, 75,1% and 69,2% of the students considered it was important for them to obtain a good score in the game. This means a high percentage of the students were trying to stay quiet or, at least, lower their voices. That is perhaps one of the reasons why the average amount of sound detected by the microphone greatly decreased in both these classes. In class C more than half the students were interested in a good score and that seems to have contributed to a decrease of the average amount of sound detected by the microphone during the use of the game. Class D is the class with the lowest percentage of students interested in a good score. If there was a way to increase class D's interest in a good score, perhaps the average amount of sound detected by the microphone would have decreased more.

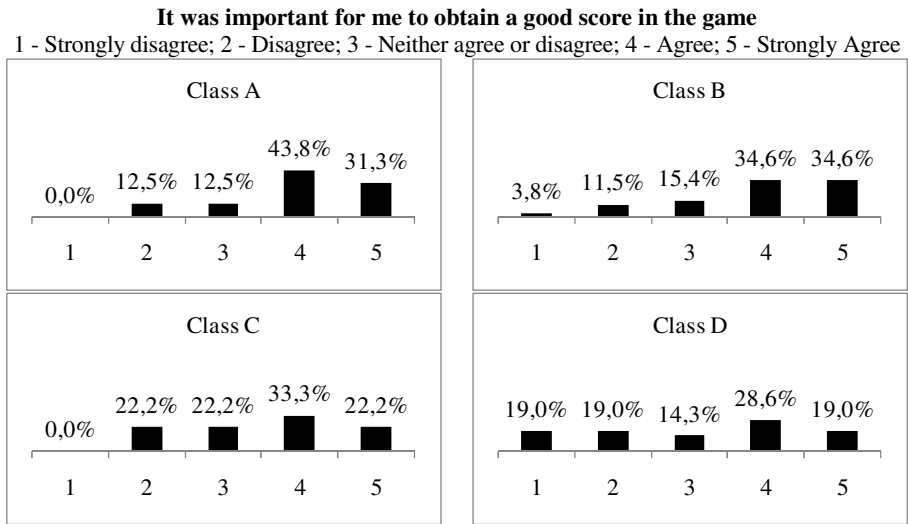


Fig. 14. Was it important for students to obtain a good score in the game?

We were also interested in knowing if the students liked the game's characters (Fig. 15). If the students liked the game's characters then maybe they could create some sort of empathy with the characters and that would motivate the students to please the character by staying quiet or lowering their voice level. In classes A and B most of the students agree or strongly agree that they liked the game's characters. In class A, the percentage of students that strongly liked the game's characters is higher than in class B, though. In class C 44,4% of the students agreed or strongly agreed that they liked the game's characters. In class D we obtained the worst results. 35% of

the students strongly disliked the game's characters and 10% didn't like them. Many of these students complained, in the survey, that the characters were too childish and, therefore, not appropriate for their age. Indeed class D is the class where the age average is higher. This seems to indicate that a different approach should have been used with these older students. Perhaps, if the students liked the characters they would feel more inclined to please them and stay quiet.

I liked the game's characters

1 - Strongly disagree; 2 - Disagree; 3 - Neither agree or disagree; 4 - Agree; 5 - Strongly Agree

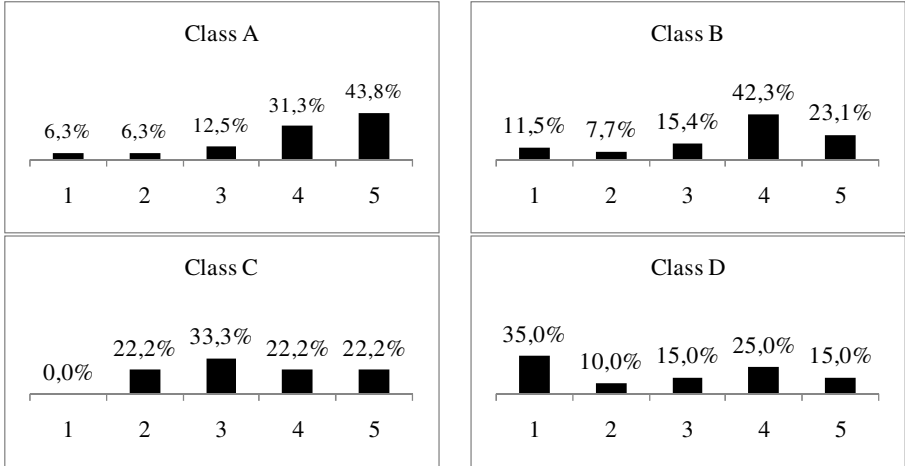


Fig. 15. Did students like the game's characters?

My favorite character is

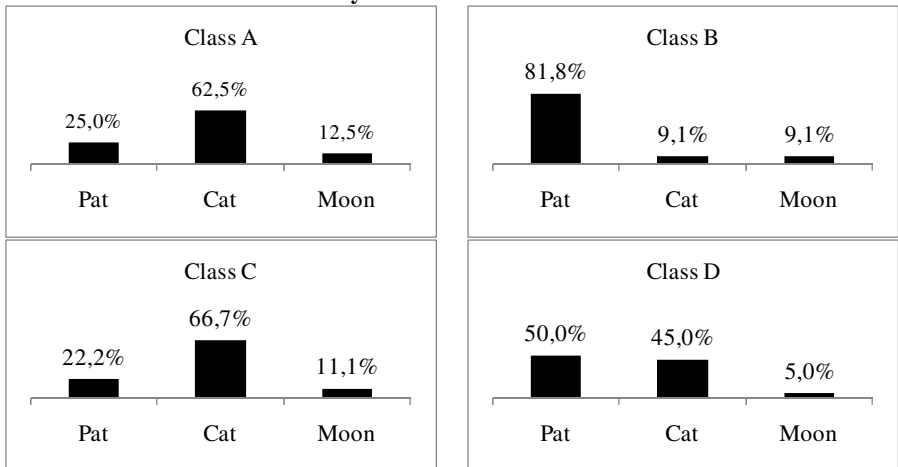


Fig. 16. Students' favorite characters

In classes A and C the favorite character was the Cat (Fig. 16). Some students told us that they found funny how the sound spectrum was integrated in the Cat's fur and how the Cat stood up when she was angry. In classes B and D the favorite character was Count Pat. The Moon was the least liked character in all the classes. Some students found the Moon too dark and dull. Others said that with Pat and with the Cat it was easier to perceive when the character was happy or angry.

7 Conclusions and Future Work

In this paper we investigated if the quantitative and graphical perception of sound, by students, in a classroom, can affect how noisy these students are. A survey was conducted to 60 teachers, of an elementary and secondary school, to find out if they considered that noise, caused by the students talking to each other in the classroom, was a problem. 78,3% of the inquired teachers agreed or strongly agreed that students make too much noise and that negatively affects their learning. Besides, 98,3% of the inquired teachers thought that when the students are making too much noise it is more difficult to teach the class. Only 34,5% of the inquired 81 students agreed or strongly agreed that their colleagues make too much noise in the classroom.

The teacher can tell the students that they are speaking too loud, but this does not show then, in a quantitative way, how much noise they are making. Our hypothesis was that, if students are more aware of how much noise they are making this would cause them to be quieter. To test this hypothesis we developed a game that shows students, in real time, the amount of sound a microphone is detecting in the classroom. The lower the amount of sound the microphone is detecting, the more points the students accumulate. The game shows the sound wave or the frequency spectrum, after a Fourier transform is performed on the sound data. The interface also integrates a character and a volume bar. The character changes state and the volume bar changes size according to the microphone's activity level. A video projector was used so that all students could see the output of the game.

The game was tested in classes A, B, C and D. The average amount of sound detected by the microphone was computed during a first lesson, for each class, without the game. Afterwards, during a second lesson, the average amount of sound detected by the microphone was computed again while using the game. The average amount of sound detected by the microphone decreased 39,6%, 26,2%, 21,2% and 5,3% in, respectively, classes A, B, C and D. The age average of classes A, B, C and D is 14,6 years old, 14 years old, 15,5 years and 17 years old respectively. Therefore, the game performed better, in the reduction of the average amount of sound detected by the microphone, with the younger students. Future work will have to be conducted to test if showing the students the amount of sound detected by the microphone in the classroom is a strategy that works better with younger students. Several students in class D complained that the game's characters were too childish. If the game's characters had been more adequate to these students then perhaps they would have felt more inclined to be quieter or to lower their voices. Nevertheless, even in class D, the average amount of sound detected by the microphone decreased 5,3%.

The influence of the game in the students' attention to what the teacher was explaining was also tested. We interviewed the teachers of the classes and conducted

an anonymous survey to the students. The answers show that the game did not negatively influence the attention of the students in class.

Acknowledgements. This work was partly funded by FCT/MCTES, through grant SFRH/BD/61085/2009, and by Centro de Informática e Tecnologias da Informação (CITI/FCT/UNL)-2011-2012 through grant PEst-OE/EEI/UI0527/2011. The authors thank everyone at IMG-CITI.

References

1. Nelson, P., Kohnert, K., Sabur, S.: Classroom Noise and Children Learning Through a Second Language. In: *Language, Speech, and Hearing Services in Schools*, vol. 36, pp. 219–229 (2005)
2. Shield, B., Dockrell, J.: The Effects of noise on the attainments and cognitive performance of primary school children: executive summary. Report for the UK's Department of Health (2002)
3. Walinder, R., Gunnarsson, K., Runeson, R., Smedje, G.: Physiological and psychological stress reactions in relation to classroom noise. *Scandinavian Journal of Work, Environment & Health* 33(4), 260–266 (2007)
4. Enmarker, I., Boman, E.: Noise annoyance responses of middle school pupils and teachers. *Journal of Environmental Psychology* 24(4), 527–536 (2004)
5. Arends, R.I.: *Aprender a Ensinar*. Editora McGraw-Hill de Portugal, Amadora (1995)
6. Alvear, R.M.B., Martínez-Arquero, G., Barón, F.J., Hernández-Mendo, A.: An Interdisciplinary Approach to Teachers' Voice Disorders and Psychosocial Working Conditions. *International Journal of Phoniatics, Speech Therapy and Communications Pathology* 62(1-2) (2010)
7. Roy, N., Merrill, R.M., Thibeault, S., Gray, S.D., Smith, E.M.: Voice Disorders in Teachers and the General Population - Effects on Work Performance, Attendance, and Future Career Choices. *Journal of Speech, Language, and Hearing Research* 47(3), 542–551 (2004)
8. Nèrière, E., Vercambre, M., Gilbert, F., Kovess-Masféty, V.: Voice disorders and mental health in teachers: a cross-sectional nationwide study. *BMC Public Health* 9, 370 (2009)
9. Gassull, C., Casanova, C., Botey, Q., Amador, M.: The Impact of the Reactivity to Stress in Teachers with Voice Problems. *International Journal of Phoniatics, Speech Therapy and Communications Pathology* 62(1-2), 35–39 (2010)
10. Hall, L., Jones, S., Paiva, A., Aylett, R.: FearNot!: providing children with strategies to cope with bullying. In: *Proceedings of the 8th International Conference on Interaction Design and Children*, pp. 276–277. ACM, New York (2009)
11. Lobo, P., Romão, T., Dias, E.A., Danado, J.C.: A Framework to Develop Persuasive Smart Environments. In: Tscheligi, M., Ruyter, B., Markopoulos, P., Wichert, R., Mirlacher, T., Meschterjakov, A., Reitberger, W. (eds.) *Proceedings of the European Conference on Ambient Intelligence*, pp. 225–234. Springer, Heidelberg (2009)
12. Bottle Bank Arcade Machine, <http://www.thefuntheory.com/bottle-bank-arcade-machine>
13. Gustafsson, A., Bång, M., Svahn, M.: Power explorer: a casual game style for encouraging long term behavior change among teenagers. In: *Proceedings of the International Conference on Advances in Computer Entertainment*, pp. 182–189. ACM, New York (2009)

14. Lin, T., Chang, K., Liu, S., Chu, H.: A persuasive game to encourage healthy dietary behaviors of young children. *Demo Paper & Adjunct Proceedings of the 8th International Conference on Ubiquitous Computing* (2006)
15. Soler, C., Zacarias, A., Lucero, A.: Molaropolis: a mobile persuasive game to raise oral health and dental hygiene awareness. In: *Proceedings of the International Conference on Advances in Computer Entertainment Technology*, pp. 388–391. ACM, New York (2009)
16. Bekker, T., Sturm, J., Eggen, B.: Designing playful interactions for social interaction and physical play. *Personal and Ubiquitous Computing* 14(5), 385–396 (2010)
17. Khaled, R., Barr, P., Biddle, R., Fischer, R., Noble, J.: Game design strategies for collectivist persuasion. In: Spencer, S.N. (ed.) *Proceedings of the 2009 ACM SIGGRAPH Symposium on Video Games*, pp. 31–38. ACM, New York (2009)
18. Oliveira, R., Cherubini, M., Oliver, N.: MoviPill: improving medication compliance for elders using a mobile persuasive social game. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pp. 251–260. ACM, New York (2010)
19. Rosenbaum, E., Klopfer, E., Boughner, B., Rosenheck, L.: Engaging students in science controversy through an augmented reality role-playing game. In: Chinn, C.A., Erkens, G., Puntambekar, S. (eds.) *Proceedings of the 8th International Conference on Computer Supported Collaborative Learning*, pp. 612–614. International Society of the Learning Sciences (2007)
20. Gamberini, L., Breda, L., Grassi, A.: VIDEODOPE: applying persuasive technology to improve awareness of drugs abuse effects. In: Shumaker, R. (ed.) *HCII 2007. LNCS*, vol. 4563, pp. 633–641. Springer, Heidelberg (2007)
21. Hirsch, T.: Water wars: designing a civic game about water scarcity. In: *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, pp. 340–343. ACM, New York (2010)
22. Shastri, D., Fujiki, Y., Buffington, R., Tsiamyrtzis, P., Pavlidis, I.: O job can you return my mojo: improving human engagement and enjoyment in routine activities. In: *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, pp. 2491–2498. ACM, New York (2010)
23. Botella, C., Breton-López, J., Quero, S., Baños, R.M., García-Palacios, A., Zaragoza, I., Alcaniz, M.: Treating cockroach phobia using a serious game on a mobile phone and augmented reality exposure: A single case study. *Computers in Human Behavior* 27(1) (2011)
24. The World's Deepest Bin,
<http://www.thefuntheory.com/worlds-deepest-bin>
25. The Speed Camera Lottery,
<http://www.thefuntheory.com/speed-camera-lottery-0>
26. Schmidt, G.W., Ulrich, R.E.: Effects of group contingent events upon classroom noise. *Journal of Applied Behavior Analysis* 2(3), 171–179 (1969)
27. Strang, H.R., George, J.R.: Clowning around to stop clowning around: a brief report on an automated approach to monitor, record, and control classroom noise. *Journal of Applied Behavior Analysis* 8(4), 471–474 (1975)
28. Groff, J.W.: Tamperproof classroom noise alarm. United States Patent, Patent number: 4654642 (Filing date October 18, 1985), Issue date: (March 31, 1987)

Encouraging Initiative in the Classroom with Anonymous Feedback

Tony Bergstrom, Andrew Harris, and Karrie Karahalios

Department of Computer Science
University of Illinois at Urbana-Champaign
{abergst2,harris78,kkarahal}@illinois.edu

Abstract. Inspiring and maintaining student participation in large classes can be a difficult task. Students benefit from an active experience as it helps them better understand the course material. However, it's easy to stay silent. Opportunities to participate in conversation allow students to question and learn. The Fragmented Social Mirror (FSM) provides students with the ability to anonymously initiate classroom dialog with the lecturer. The system encourages participation by enabling expressive anonymous feedback to reduce evaluation anxiety. The FSM further catalyzes participation by allowing for many simultaneous participants. In this paper, we introduce the FSM as a classroom device, discuss its design, and describe a pilot test of the interface. Initial results indicate a promising direction for future feedback systems.

Keywords: Social Mirrors, Classroom, Feedback, Anonymous.

1 Introduction

Students learn more when they actively engage in the classroom [22]. However the structure of many classes ensures that the lecturer speaks for at least 80% of the time. Though some students participate, it's expected that five students out of 40 will come to dominate any classroom discussion.

The lecturer's awareness of class comprehension is skewed both by the students' many social pressures and the few speaking opportunities. Students try to present a positive image of themselves to their peers. Thus, they often avoid volunteering information due to evaluation anxiety, a fear of being judged by others for making a mistake or being the focus of attention [22]. It's easy to remain silent. Those students who do speak are generally self-confident or understand the material. However, there is a reluctance to appear *too* engaged in the classroom. Students who raise the expectations on a group may be ostracized by their peers [17].

In this paper, we present the design of an interface prototype to encourage student engagement and improve the lecturer's awareness in the classroom. The prototype, entitled Fragmented Social Mirror (FSM), aims to create a new communication channel of anonymous dialog between the instructor and the class. Unlike many previous Audience Response Systems [10, 19], FSM allows for expressive text-based feedback and may be used throughout the lecture. In our short pilot observation, students in a large class began to initiate interaction with the professor, whereas previously they

had only mumbled answers in response to posed questions. In this paper, we describe the design of the FSM in the context of other Audience Response Systems. We also discuss promising initial observations from a classroom pilot study of the FSM.

2 Feedback in Conversation

FSM is designed to extend the benefit of backchannel communication. Familiar face-to-face backchannels include “yeahs,” “uh huhs,” and head nods that show attention to the speaker. Similarly, facial expressions reveal feelings while gestures provide emphasis [15]. From the listener’s perspective, these signals show the speaker that the audience is listening and is interested. In a large audience, these visual signals can be lost in the crowd, which necessitates speaking up, murmuring, and applause. Large lectures and speeches require a strong and focused mediation to centralize attention [8]. Without that centralization, small groups of 4-5 individuals form based on proximity of others.

The classroom environment has to balance the need for a strongly focused discussion with the need for student feedback. Lack of participation makes it more difficult to assess the current understanding of students and by not encouraging participation students are less apt to risk being wrong [12]. However, from a constructivist view of learning theory, students need to be actively engaged in their learning. Students learn by actively building their own understanding of new information [1, 21]. Instruction alone does not directly allow for the construction of knowledge, however effective knowledge construction often adopts a social process between the student and the teacher [7].

Classrooms have addressed this issue by using Audience Response Systems for multiple choice and true/false questions. The interfaces in [10, 19] provide a small number of preselected responses of A/B/C/D, and a true/false response. These interfaces are most often used when the lecturer explicitly asks a multiple-choice question of the audience. To be effective, the lecturer has the difficult task of anticipating key moments to query the audience and must specifically structure a lecture to accommodate this new question-answer format. Though each system varies, many include specialized hardware, which are either purchased by or provided for each student [14, 19]. In the worst cases, when a lecturer does not incorporate interesting interaction into the lecture, the Audience Response Systems can become automated attendance and quiz systems, which students grow to resent [14].

Other feedback modalities such as text-based systems provide opportunities for students to engage with each other [16, 23]. Studies of dedicated course chatrooms show students will chat about the lecture’s content to help explain concepts to confused classmates [23]. However, chat rooms also encourage unrelated discussions, and potentially draw students away from the lecture. Some have gone further and tied the in-class chat discussion to a video recording of the lecture for archival and review [3]. For a practice talk or presentation, these systems allow people to access the initial thoughts of the audience in an asynchronous manner. In addition to making help available, this style of active learning helps students communicate concepts to peers for a deeper understanding of the material. Others have brought affective computing to the classroom by using a custom handheld device [4]. This handheld ball can be used to

indicate the emotional state of the student to the instructor. Outside of the classroom, text based systems similarly open public dialog into shared events through IRC, instant messaging, Twitter, and Facebook [9, 16, 18]. Though all of these side-channels can contribute to audience discussion, they often leave the speaker out of the loop during the event.

Writing systems, such as Classroom Presenter, benefit from writing as input while still including the instructor in the interaction [2]. A tablet PC system, it allows students to mark directly on the current slide with a stylus, which can then be viewed and shared by students and the instructor. Instructors can set up slides that encourage students to answer questions that can be discussed and reviewed as a group. It also allows for more expressive diagrams, images, and nonverbal communication mechanisms. This method enables a broad sampling of student understanding and encourages active participation with the material.

Related work such as backchan.nl and Conversation Votes create a new feedback channel that integrates approval feedback into group dialog [6, 11]. With backchan.nl audience members organize their collective questions for the speaker in a conference or after a talk. A moderator filters the most appropriate questions from the top rated questions. With Conversation Votes, participants annotated an abstract visualization of conversation with positive and negative votes to highlight agreement during conversation. In small groups, this anonymous feedback increased the level of participation from those less satisfied with previous conversations.

Viewed on an axis of expressivity, distinct categories of low expressivity and high expressivity emerge. Low expressivity systems as in [6, 10, 19] limit what a student can communicate, but ensure the feedback can be quickly interpreted. High expressivity systems like [2, 11] and chatrooms allow students vast communication capabilities, but can require more focused attention for both the lecturer and students.

Our work takes a middle path. FSM provides a meaningful, but constrained, set of signals to be observed alongside the lecture like low expressivity systems, but it allows expressive text to convey personal ideas as in a high expressivity system. As an always-available interface, the FSM captures the fleeting moments of confusion and conveys this information to the lecturer while it can be addressed in context.

3 The Fragmented Social Mirror

The Fragmented Social Mirror (FSM) provides feedback based on principles borrowed from our previous work in social mirrors [13]; however, the classroom setting necessitates a break from the standard social mirror design. A social mirror is a real-time depiction of interaction meant to augment natural face-to-face environment. It captures ephemeral moments in conversation and brings them into the public view through visualization. In our previous work, social mirrors displayed abstract visualizations to depict participation in conversation. The resulting display of conversational dominance, non-participation, and turn taking encouraged more balanced conversation [6]. In these social mirrors, *one* shared visualization of conversation was projected centrally for all participants to see. On a classroom or large lecture scale, this form does not function as well. There are many more participants involved, and the architecture of the space is different from the spatial layout of small group interaction

around one shared table. Furthermore, there is a natural asymmetry in participation due to the lecturer-audience dynamic. This results in less interaction between the lecturer and the audience not suited to the traditional social mirror visualizations.

The term “fragmented” in FSM refers to the use of individual interfaces for each participant as opposed to one shared visualization and the shortened time component as opposed to the full history present in previous social mirrors. In our setup, each participant accesses a Java applet from his or her computer or mobile device, while a large public display is presented to everyone. Furthermore, while a traditional social mirror maintains a persistent history of interaction, FSM highlights questions and comments that are pressing at the specific moment.

3.1 FSM Design Choices

The FSM design focuses on capturing and reflecting the unheard and unvoiced dialog in the classroom. The current design is the result of a long chain of prototypes that sought to balance the need for attention to the interface with attention to the classroom. The final design in this paper served as a culmination of this prototyping, though our pilot demonstrated areas for further refinement such as moderation.

The process of designing the FSM began by observing an active and engaged classroom of 100+ students to see what students say when engaged in an active class. The lecturers of these classes were generally rated as among the best in the department. They were engaging during lecture and good at encouraging student participation. To facilitate more participation the lecturers posed a question and waited for responses - thus guaranteeing an answer or a question for clarification. We noted all the types of student responses to better understand what a student wants to say during class. The responses were narrowed down to the following list of categorical responses:

- **Questions:** Students provide new questions based on what has just been taught.
- **Information:** Students add their own connection to outside subjects.
- **Agreement/Disagreement:** Answering a Lecturer’s question.
- **Slow Down/Redo:** Students did not understand the lecturer.
- **Cannot Hear/Repeat:** Students did not hear the lecturer.

Our list was very similar to feedback available in other work to mark up a presentation slide [20]. We began to investigate this set of six messages for our prototypes. These six messages would serve as categories with the ability for students to include a short text message for explanation. The message categories serve as a means to identify and group similar responses and highlight important categories like questions. In parallel with our interface design, we investigated imagery for each of these six categories of messages (described in the next section). Due to this process, “Slow Down/Redo” and “Cannot Hear/Repeat” were eliminated. Suitable icons could not be found and they easily be replaced by an “Information” message with appropriate text.

Many of our initial interface prototypes borrowed design components from the *Conversation Clock* and *Conversation Votes* [5, 6], they incorporated the feedback into a timeline that structured the activity throughout the session. In some cases, we included indications of speaker. Much like a standard instant messenger, the full

history of messages could be read through at any time. These interfaces showed potential for the review of archival classroom data, but did not serve our purpose of encouraging classroom interaction. These prototypes, tested amongst our own group, required too much attention to adequately understand.

After refining the initial prototypes, we settled on a simple interface students could use without pulling their attention too far from the lecturer. Thus the input of the FSM was used only for capturing one comment. The history of feedback was only seen on the public display and limited that history to the most recent comments. Additionally, the needs of the lecturer necessitated this type of design. The lecturer needed to be able to read feedback from the hundreds in the audience while still being able to teach effectively. In past studies, a social mirror was primarily viewed by the listeners (and not the speaker) in conversation because they had more free attention [5]. In this design, the captured feedback of conversation is significantly pared down, so that the lecturer can receive the benefits from the social mirror with minimal attention. Therefore, current comments/questions are displayed so as not to overwhelm the viewers with a long history.

Iconographic Messages. The FSM interface passes information through icons. These graphics serve to simplify the message so that the lecturer might easily understand the classroom without reading too much content. Based on informal observation of classroom sessions and prior work [20], we designed icons based on the messages earlier: “I have a question,” “I have information/an answer,” “Yes/agree,” “No/disagree,” “Speak Up,” “Slow Down.” Three researchers independently drew any graphic that they felt reasonably captured these messages. We combined them into sets for each category, with a total of 5–15 images for each message.

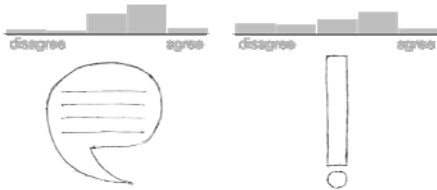
We conducted a survey of Computer Science undergraduates to test our icon designs. A total of 54 Computer Science undergraduates completed our survey. Their feedback identified 17 icons that convey the intended message. Figure 1 shows all 17 icons. None of the icons for “Slow Down” conveyed an adequate message to the student. We eliminated this message, as well as the “Speak Up” messages in favor of a simpler 4-icon interface. Students can use the Information and Question messages with additional text to signal “Slow Down” and “Speak Up.”

FSM Interfaces. There are two FSM interfaces — the student’s client interface for a computer or handheld device (Figure 2) and a larger public screen for the lecturer and audience (Figure 3). The public display is situated in the front of the room, though the lecturer sees the public display on a personal screen. The four different preselected icons categorize student responses in the student interface. The icons represent: Information, Questions, yes/agree, no/disagree. Of the four categories or signals, the Information and Question signals can be augmented by a 40-character message. The short messages allow students to clarify their questions or possible answers when there is no opportunity to speak while the yes/no buttons allows students to answer simple questions quickly.

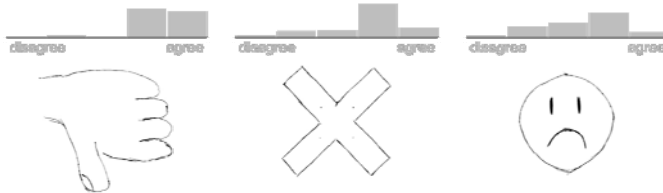
The icon represents confusion or the desire to ask a question.



The icon represents an audience member has new information.



The icon represents disagreement or a negative response.



The icon represents agreement or a positive response.



The icon represents that the lecturer needs to increase their volume.

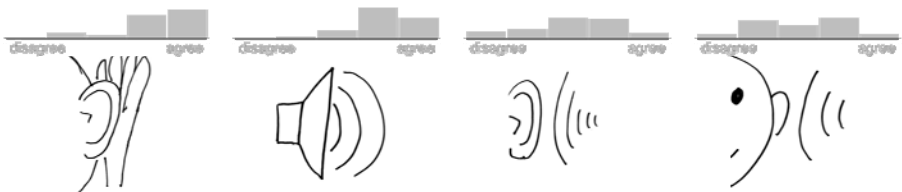


Fig. 1. The results of 54 individuals rating our sets of icons produced these icons as the most indicative of their intended messages. The image survey informed our final selection of icons seen in the interface (Figures 1 and 3). One other message, to indicate the speaker was moving too fast, produced no positively rated icons.



Fig. 2. The input device is small and simple for classroom use. The two left icons for information and questions allow for typing phrases to send along with the icon message.

Students use the client interface in Figure 2 to send their message to the public display shown in Figure 3. All messages on the public display are grouped by their associated icon to increase legibility for the speaker. The speaker can look up and see many questions that need to be addressed or they can glance over answers that students provided via the display. The icon group with the most messages moves to the top of the screen with a larger icon. The most recent message of this icon appears at the top of that icon in white text set against the black background. As a message ages, it fades to grey before finally disappearing after a pre-configured time. For icons with multiple messages, a count is displayed to the left of the icon.

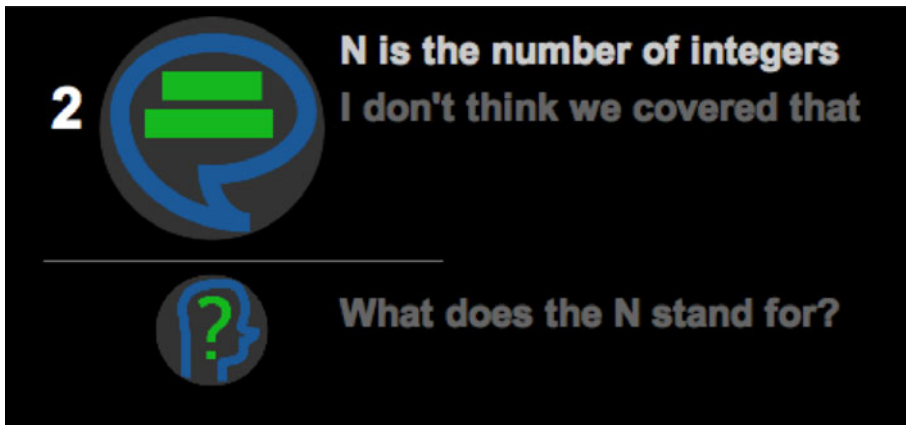


Fig. 3. The public display groups messages by icon and highlights the most recent feedback from the class. To the left of each icon, a counter indicates multiple messages of the same type, particularly useful when using the yes/no feedback buttons.

Messages on the public display are limited to recent messages. Only the most recent minute of activity is visible; each message fades in brightness over the minute before disappearing from view. The rationale for this design was two fold: (1) we did not want the lecturers to be confused or overwhelmed by reading old questions from a prior part of the lecture and (2) if a question goes unanswered and disappears, this

removal may encourage a student to verbalize the question in class or to repost it. One of our main goals is to encourage more class interaction. If a student can “see” that they are not alone in their confusion, they may be less apprehensive to speak out and ask a question.

Once a student sends a signal via posting an icon, they are blocked from sending additional signals for a brief period (10 seconds in our pilot) to discourage excessive social chatter and monopolization of the channel. While there is some room for abuse as with the backchan.nl system, where some users voted up questions for humor [11], the public availability of the channel is ultimately at the discretion of the lecturer.

4 Pilot Study

We conducted a pilot study to investigate the FSM in the classroom. We began by observing the participation levels before the introduction of the FSM and again with the FSM in place. For this, we observed a required second year course with roughly 180 registered students at the beginning of the semester. The instructor was not affiliated with our research team. We observed a total of six course sessions: three initially without any augmentation, and three with the addition of the FSM. During observation, an average of 100.0 students were in attendance, though there were fewer students in the final sessions (attributed to an intervening midterm and final day to drop the course). Given the large class, not many students had the opportunity to speak, and most did not. A summary of the attendance is visible in Figure 4.

Session	Students	Computers	On FSM
1	108	17	
2	125	25	
3	112	19	
4	102	23	14
5	80	19	15
6	73	14	5

Fig. 4. This table shows the number of people in the classroom for each session as well as the number of computer visible. Though we did request students to bring their computers to the sessions with the Fragmented Social Mirror. The number of computers remained essentially unchanged.

Prior to testing the FSM in class, we sent a pre-survey and described the use of the FSM. The survey inquired about the student's comfort level while participating in class versus their smaller discussion sections. Feedback from the survey confirmed that students are not comfortable asking questions or asking for clarification during class, though they are more comfortable asking in their smaller recitation sections. Similarly, they recognize that they do not participate or ask questions during class (Figure 5).

Preliminary Survey Responses and Questions

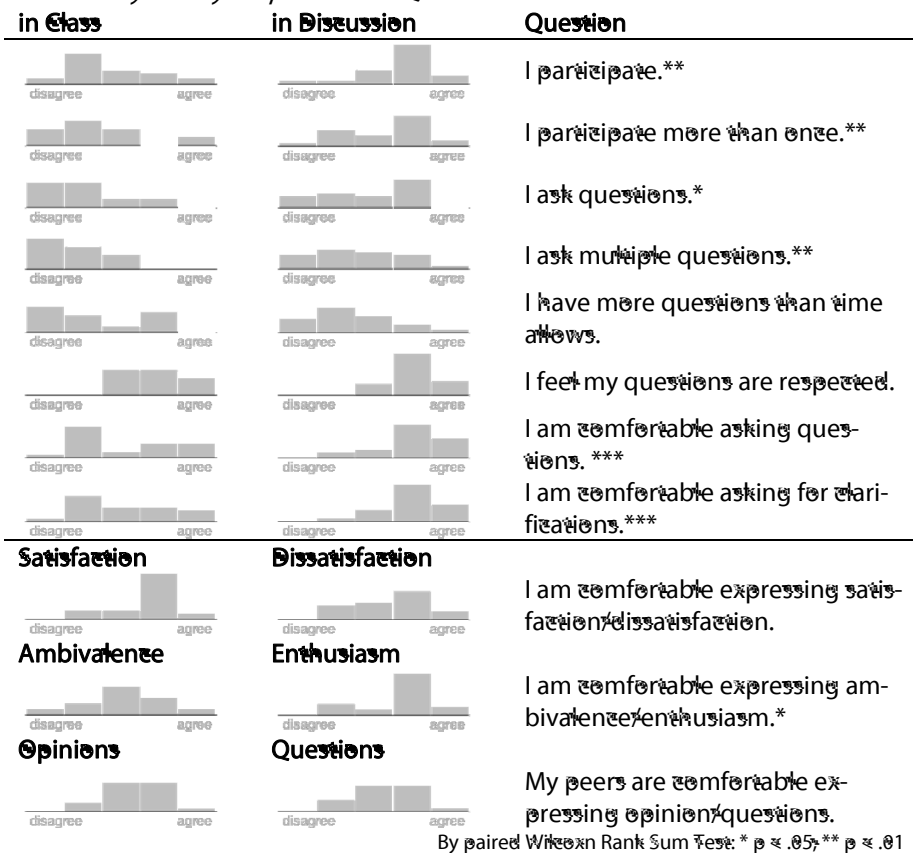


Fig. 5. Students reported they were uncomfortable asking questions in class, though it was less the case in smaller recitation sections. They are aware that they do not participate in class and are less likely to ask questions even though they have them. There is reluctance amongst the students to speak up and participate. Survey with 23 respondents. All graphs have the same scale.

Our initial observations showed little interaction between audience and lecturer over the course of three 50-minute sessions. The only activity from the audience was in response to questions posed by the lecturer. For example, students were asked "n is divisible by what?" and "What is the cardinality of set Q?" in reference to a proof. The class averaged about four responses per class. The students initiated zero interactions themselves, five of the twelve responses were general indefinite murmurs from the class, and two responses involved raising hands. Various sets of 1–3 unidentified students spoke up to answer the remaining six questions.

We tested the FSM in three class sessions and found the students were proactive in using the system. In the classroom, the lecturer used a central projection screen to work through problems by hand while a smaller screen displayed the public display to the right of the larger screen (Figure 6). At the lecture podium, the lecturer also had a copy of the public display available during the class activity.

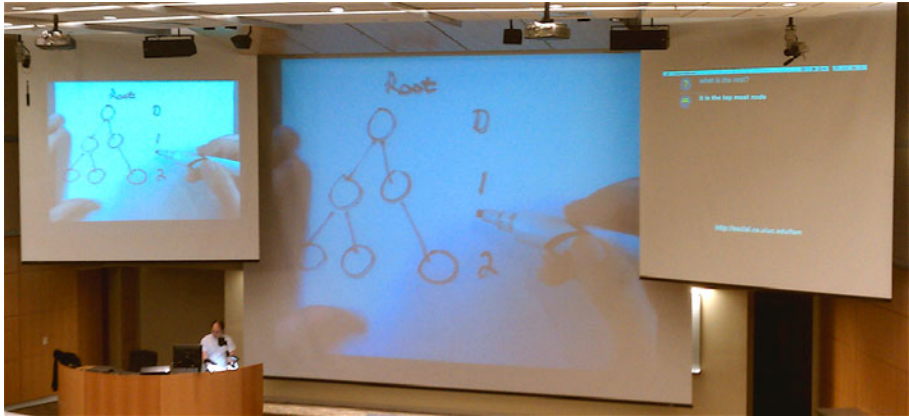


Fig. 6. We tested the Fragmented Social Mirror in a large lecture hall with three projection screens. Typically, the instructor repeated the same material on all three screens. For the study, the Fragmented Social Mirror replaced one screen during class time. The screen by the lecturer also displayed the public interface.

With the system in place, Students initiated dialog with the lecturer by asking questions 11 times, compared to zero without the system. When on topic, students used the Fragmented Social Mirror to ask questions of the professor, keep the professor from moving on too quickly, and to answer any questions the professor posed. Figure 7 summarizes the participation in each of the 6 classes. Most of the on-topic dialogs either began with or contained a question for the instructor. They lead to discussions with the instructor and information to enrich the class. However, there were also many off-topic messages. These messages were irrelevant to the class topic and were used to draw the attention of other classmates away from the lecture material for their own entertainment.

Session	Questions Posed by		Dialog Relevance	
	Instructor	Student	On-topic	Off-topic
1 No FSM	3	0	3	0
2 No FSM	3	0	3	0
3 No FSM	4	0	4	0
4 FSM	2	2	6 (22)	2 (2)
5 FSM	2	7	10 (30)	4 (37)
6 FSM	1	2	6 (8)	3 (7)

Fig. 7. The Fragmented Social Mirror encouraged students to initiate questions and dialog with the instructor. Many dialogs began with or contained questions, though some were responses to indicate comprehension of the material. The anonymity of the interface also encourages unrelated dialog in class. A dialog in the Fragmented Social Mirror could include multiple messages – thus, # dialogs (# messages) includes both the count of interaction instances and total messages.

Example FSM Dialogs. Excerpts from the FSM sessions appear below. In this first example, students requested information that the professor was not trying to teach but established an interesting aside on history related to the lesson:

Instructor [discussing the Karatsuba Algorithm]



<What is Karatsuba?>

Instructor Karatsuba is the guy who invented it, Anatolii Karatsuba. [Instructor continues with a bit more history.]

As another example, the student's lack of understanding prompts him or her to ask for clarification on calculating tree height.

Instructor [Providing an explanation of tree depth]



wow



What is the height again?



The maximum depth of the tree. You can count the Levels by generation



Not 5? It's max level and not count?

Instructor Yeah, it's 4 not 5 ... [continues on 0 based counting]



In cs you start counting at zero :)

However, with the addition of initiating comments, there was also an increase in comments solely intended to draw attention away from lecture. These messages often had nothing to do with the lecture or a question tended to come in bursts in order to overwhelm the public display for a short time. As an example of such a burst:



HATE HATE HATE HATE HATE HAT



I DONT CARE WHAT THESE CHICKS SA



I DONT EVEN LOOK THAT WAY



EVERY TIME I WALK IN THE CLU



THEY HATIN ON ME CUSE THEY KNO I LK GOO

This type of interaction was most prevalent in the second session. The lecturer was inclined to read them, see that they were not relevant and either laugh, if it were funny, or state “I don’t know what this means.” However, the increase of messages also meant that the lecturer was more likely to miss relevant exchanges where a student was asking for help:



can you draw the picture for the tree afte



after applying rule 3



:(



+1

After the Sessions. We had only planned to gather initial observations to refine the system in these first sessions; however, the instructor was excited to see the students participating and invited us to return with the system for further studies. After the lectures, she indicated that it’s always been hard to get this many students to say anything, even with encouragement. The simplicity of the display was also deemed useful, as she could read the questions with a glance. Additionally, the asynchronous nature allowed students to ask their questions while she was still explaining — thus allowing her to work the question into that explanation or come back to it later. Student feedback indicated the device was useful as they “didn’t have to try to get the

professors attention” by raising a hand from the back of the lecture.

Students also saw the benefit of the interface, and felt it was easier to participate in the classroom (Figure 8). However, they recognized the difficulty of maintaining order in the anonymous display and provided suggestions to keep the interface on topic. One such suggestion was to make the display semi-anonymous; implement a publicly anonymous interface that retains the identity on the lecturer's display. In this way the instructor could call out any abuse of the display, while protecting the identity of any others who were uncomfortable commenting in front of the class. A similar suggestion would simply log the identities for review after class.

5 Discussion

The Fragmented Social mirror touched the surface of integrated feedback in large discussions by allowing a controlled set of feedback that could express a student's question or response. As we mentioned in related work, the field can be divided into the high expressivity interfaces that require more attention and the low expressivity interfaces that do not allow students to indicate their questions.

In the sessions with the FSM, the classroom dialog was more involved. The lecturer felt like she was talking to people rather than at people while the students took a more proactive role in directing conversation to points that were not understandable. With 100 students, evaluation anxiety limits the individuals willing to speak - however we have shown that anonymous feedback can break the barrier and include more students.

Our interface was not perfect; the classroom sessions revealed that while anonymity opens opportunities for discussion, it must be tempered in some manner. In both the second and third sessions, some individuals engaged in the public conversation by adding potentially disruptive comments that lead to legitimate questions not being discussed. As a prototype, we did not fully flesh out any mechanisms to prevent this type of interaction. Perhaps the lecturer should be able to flag such comments during conversation to lock out individuals, or the lecturer should be able to identify individuals after the classroom session and deduct points in some manner. While it is tempting to simply allow the class to moderate itself, the design must be careful not to become more of a distraction as it requires interaction outside the scope of a learning task and draws them away from the lecture.

The classroom sessions also made it apparent that the positive and negative responses should be redesigned to provide more flexibility. Though students did use them as feedback for the lecturer, they also use the agreement checkmark to indicate a “me too” when other students raised a question. Others adopted a convention of adding a “+1” as seen in the example conversations. A revamped system might allow a student to indicate “me too” and ensure the question stays visible long enough for the lecturer to see the question.

Our surveys underscore the need for large classrooms to tap into technological backchannels. Students know that they do not participate in large classroom settings even though they have questions. They are not comfortable asking questions in such a large group. The survey after the use of our system shows that students felt encouraged themselves and the class to ask questions, they found it to make the lectures more enjoyable, and it was a worthwhile addition to the lecture.

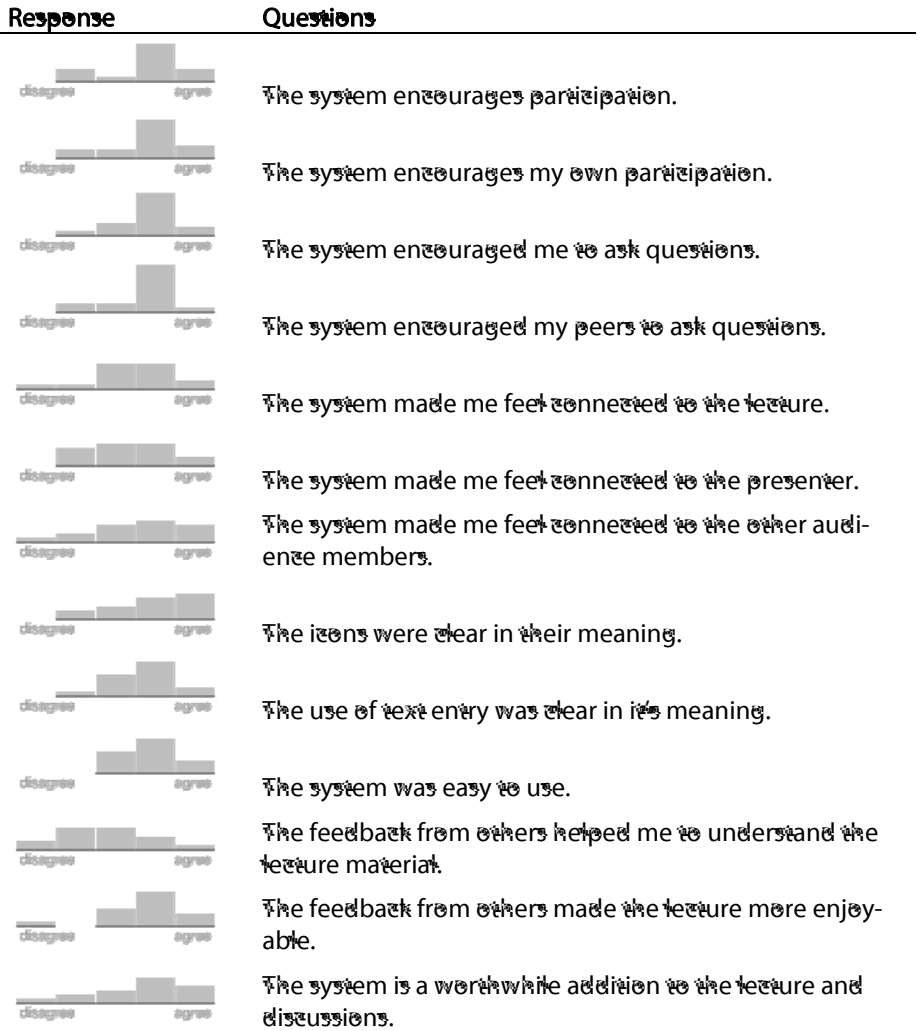


Fig. 8. Students reported the Fragmented Social Mirror encouraged participation and connected them to the lecture. Students reported the lectures being more enjoyable and saw the feedback as a worthwhile addition to the lectures.

The FSM interface received a positive response from both students and lecturer. Our initial study highlights the use of anonymous signals in large classroom has potential to draw in more active participation of the students and audience.

6 Conclusion and Future Work

A large audience automatically reduces the possibilities for participation in a lecture setting. Rather than accept this and move on, technology has provided new channels to engage students throughout the lecture.

Our own work shows that expressive feedback can be gathered from lightweight systems. The use of these channels was shown to increase engagement, to benefit the students, and to spark initiative where there was none before.

While our system was successful in engaging students and encouraging participation, we acknowledge it is not a definitive study and has limitations. Our study was a small study to test our conception of anonymous feedback. The system, with refinements, should be further tested over a longer term and in multiple classrooms. Many untested facets of the interface can be further explored. We advocate anonymous feedback based on the premise of evaluation anxiety, though we have not yet tested the effects of allowing or enforcing identity in the FSM.

The Fragmented Social Mirror indicates that the use of text based anonymous feedback has potential for promoting engagement in the classroom. A long term study could investigate the effects on learning outcomes: does the FSM encourage students who are already engaged in class to further surpass their peers, or does it genuinely help students who just need a small boost to get involved?

We hope to explore these further questions on the benefits of classroom feedback in future works with the FSM and other interfaces.

References

1. Anderson, C.W.: Strategic teaching in science. In: Jones, B.F., Palincsar, A.S., Ogle, D.S., Carr, E.G. (eds.) *Strategic Teaching and Learning: Cognitive Instruction in the Content Areas* (1987)
2. Anderson, R., Anderson, R., Davis, P., Linnell, N., Prince, C., Razmov, V., Videon, F.: Classroom Presenter: Enhancing Interactive Education with Digital Ink. *Computer* 40(9), 56–61 (2007)
3. Baecker, R., Fono, D., Lillian, B., Collins, C.: Webcasting made interactive: persistent chat for text dialogue during and about learning events. In: Smith, M.J., Salvendy, G. (eds.) *HCI 2007, Part II. LNCS*, vol. 4558, pp. 260–268. Springer, Heidelberg (2007)
4. Balaam, M., Fitzpatrick, G., Good, J., Luckin, R.: Exploring affective technologies for the classroom with the subtle stone. In: *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010* (2010)
5. Bergstrom, T., Karahalios, K.: Seeing More: Visualizing Audio Cues. In: Baranauskas, C., Abascal, J., Barbosa, S.D.J. (eds.) *INTERACT 2007. LNCS*, vol. 4663, pp. 29–42. Springer, Heidelberg (2007)
6. Bergstrom, T., Karahalios, K.: Vote and Be Heard: Adding Back-Channel Cues to Social Mirrors. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) *INTERACT 2009. LNCS*, vol. 5726, pp. 546–559. Springer, Heidelberg (2009)
7. Brown, J., Collins, A.M., Duguid, P.: *Situated cognition and the culture of learning* (1989)
8. Dunbar, R.I.M., Duncan, N.D.C., Nettle, D.: Size and structure of freelyforming conversational groups. *Human Nature* 6(1) (1995)
9. Ebner, M., Reinhardt, W.: Social networking in scientific conferences - Twitter as tool for strengthen a scientific community. In: *Proceedings of the 5th EduMedia Conference* (2009)
10. Fitch, J.L.: Student feedback in the college classroom: A technology solution. *Educational Technology Research and Development* 52(1), 71–77 (2004)

11. Harry, D., Green, J., Donath, J.: Backchan.nl: integrating backchannels in physical space. In: Proc. of CHI (2009)
12. Jones, M.G., Gerig, T.M.: Silent Sixth-Grade Students: Characteristics, Achievement, and Teacher Expectations. *The Elementary School Journal* 95(2) (1994)
13. Karahalios, K., Bergstrom, T.: Social Mirrors as Social Signals: Transforming Audio into Graphics. *IEEE Computer Graphics and Applications* 29(5), 22–32 (2009)
14. Kay, R.H., LeSage, A.: Examining the benefits and challenges of using audience response systems: A review of the literature. *Comput. Educ.* 53(3), 819–827 (2009)
15. Krauss, R.M., Garlock, C.M., Bricker, P.D., McMahon, L.E.: The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology* 35(7), 523–529 (1977)
16. McCarthy, J.F., Boyd, d.m.: Digital backchannels in shared physical spaces: experiences at an academic conference CHI 2005: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1641–1644 (2005)
17. Parks, C.D., Stone, A.B.: The desire to expel unselfish members from the group. *Journal of Personality and Social Psychology* 99(2), 303–310 (2010)
18. Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweet the debates: understanding community annotation of uncollected sources. In: WSM 2009: Proceedings of the First SIGMM Workshop on Social Media, pp. 3–10 (2009)
19. Stowell, J.R., Nelson, J.M.: Benefits of Electronic Audience Response Systems on Student Participation, Learning, and Emotion. *Teaching of Psychology* 34(4), 253–258 (2007)
20. VanDeGrift, T., Wolfman, S.A., Yasuhara, K., Anderson, R.J.: Promoting Interaction in Large Classes with a Computer-Mediated Feedback System. University of Washington, Computer Science and Engineering (2002)
21. von Glasersfeld, E.: Cognition, Construction of Knowledge, and Teaching. *History, Philosophy, and Science Teaching* 80(1), 121–140 (1989)
22. Weaver, R.R., Qi, J.: Classroom Organization and Participation: College Students' Perceptions. *The Journal of Higher Education* 76(5), 570–601 (2005)
23. Yardi, S.: Whispers in the Classroom. In: McPherson, T. (ed.) *Digital Youth, Innovation, and the Unexpected* (2008)

U-Note: Capture the Class and Access It Everywhere

Sylvain Malacria¹, Thomas Pietrzak^{1,2}, Aurélien Tabard^{1,3}, and Éric Lecolinet¹

¹ Telecom ParisTech – CNRS LTCI UMR 5141, 46 rue Barrault, 75013, Paris, France
{malacria, elc}@telecom-paristech.fr

² University of Toronto, 40 St. George Street, Toronto, Ontario, Canada
pietrzak@dgp.toronto.edu

³ IT University of Copenhagen, Rued Langgaards Vej 7, DK-2300 Copenhagen, Denmark
auta@itu.dk

Abstract. We present U-Note, an augmented teaching and learning system leveraging the advantages of paper while letting teachers and pupils benefit from the richness that digital media can bring to a lecture. U-Note provides automatic linking between the notes of the pupils' notebooks and various events that occurred during the class (such as opening digital documents, changing slides, writing text on an interactive whiteboard...). Pupils can thus explore their notes in conjunction with the digital documents that were presented by the teacher during the lesson. Additionally, they can also listen to what the teacher was saying when a given note was written. Finally, they can add their own comments and documents to their notebooks to extend their lecture notes. We interviewed teachers and deployed questionnaires to identify both teachers and pupils' habits: most of the teachers use (or would like to use) digital documents in their lectures but have problems in sharing these resources with their pupils. The results of this study also show that paper remains the primary medium used for knowledge keeping, sharing and editing by the pupils. Based on these observations, we designed U-Note, which is built on three modules. U-Teach captures the context of the class: audio recordings, the whiteboard contents, together with the web pages, videos and slideshows displayed during the lesson. U-Study binds pupils' paper notes (taken with an Anoto digital pen) with the data coming from U-Teach and lets pupils access the class materials at home, through their notebooks. U-Move lets pupils browse lecture materials on their smartphone when they are not in front of a computer.

Keywords: Augmented classroom, digital pen, digital lecturing environment, capture and access, digital classroom.

1 Introduction

Many teachers are now comfortable with digital media. Additionally, digital equipment such as PCs, video projectors, interactive white boards (IWB), etc., have become increasingly affordable. This makes it possible to use digital material in the classroom, not only at the university level, but also in middle and high schools. Laptops and mobile devices are now widespread, so that most pupils can work with digital documents at home, at the library, at other pupils' places and even in public transportation.

Various studies of augmented classrooms have already been published [1,4,12,14] with a focus on colleges and universities. In this paper, we present U-Note, a system designed for middle and high schools. The needs and the requirements of universities are different from those of high schools. From our preliminary interviews, we observed the importance of handwritten notes. Pen and paper are still the main tools used by pupils in class for several good reasons, pen and paper are cheap, flexible, easy to use, and not distractive [13,15]. But, back at home when the pupils read what they wrote in their notebooks, they cannot easily access the digital media presented during the class. They also don't have any way or reviewing the teacher's oral explanations they may have missed. Finally, while pupils now commonly use computers at home and elsewhere, there is no simple way to link their digital work (for example, searching and reading web pages) with their notebooks, which remain their main means for storing, organizing and retrieving information.

We designed U-Note by taking these findings into account. U-Note aims at linking the pupils' handwritten notes with the material that was presented during the class, with a high level of granularity. The pupils' notebooks serve as a means for referencing and accessing of digital media, oral explanations and the writing on the whiteboard, hence providing a physical medium for retrieving information from various sources. The notebooks provide a link between the pupils' works in class, at home and at any other locations. Furthermore, notebooks are designed for active reading so that pupils can enrich their personal libraries by adding their own digital documents. In all cases U-Note allows fine-grained correspondence. For instance, a phrase, symbol, or drawing can be linked with a simple slide of a presentation or an excerpt located at a specific location in a web page.

This paper begins with a presentation of existing annotation and note taking systems. Next we present interviews with elementary, middle and high school teachers. This stage helped us to refine our goals and to focus on our users' needs. We then describe U-Note and the features that appeared to be useful according to our investigations. Finally we conclude and present future work.

2 Related Work

2.1 Presentation Tools

Various systems allow subsequent access to captured live experiences. With Ubiquitous Presenter [22], a classroom presentation tool, the instructor can annotate slides with a Tablet PC while showing them and giving the lecture. The students can view the live presentation with narration and digital ink using standard PCs. The captured presentation is saved to a web server and can be retrieved later by students as a video. Recap [8] enables users to capture the lecture with more details than Ubiquitous Presenter; the presentation is indexed by slide number and by the pen strokes. Students can access this capture after the class through an ActiveX enabled web browser. However, neither Ubiquitous Presenter nor Recap provides the capability to link the multimedia data shown in class with students' notes.

Classroom2000 [1], which later became eClass [4], is a classroom presentation tool that allows an instructor to annotate slides on an interactive whiteboard. These

annotations are linked with a video and audio recording of the class, and with the web links opened during the lecture. A longitudinal evaluation of this tool showed the usefulness of the links between the documents and the audio recordings. It also underlined the fact that students took fewer notes when using the system, which is not surprising since the teacher provides his or her notes. The advantage of this is that students may concentrate on the material. However, some authors have argued that taking notes has an important role in the memorization process [7].

StuPad [19], which integrates a note taking system with pen-based video tablets, provides students with the ability to personalize the capture of the lecture experiences. However, such equipment is currently not suited to middle and high school where paper notebooks are still widely used. Besides, as demonstrated in [13,15], interfaces departing from classic GUIs such as pen tablets and graphical tablets tend to deteriorate performance, especially for low-performing students.

The Digital Lecture Halls (DLH) project focuses on large audiences and provides the lecturer with a tool to control his lecture through a dedicated interface on a pen-based tablet [12,14]. The lecturer can thus write on the digital blackboard through the tablet and annotate his presentation, while keeping eye-contact with the audience. The audience can use specific software on their digital devices (such as smartphones or laptops) to mark parts of a lecture as particularly interesting, or to ask a question. While these solutions make perfect sense for large and mature audience lectures, middle school and high school have a much smaller and co-located audience where contact with the pupils is easy. Furthermore pupils tend to be easily distracted and new equipment and software that interferes with the course may overly distract pupils.

These systems combine all captured data into a unique stream and broadcast it through a web interface or as downloadable videos. The students cannot open (or eventually edit) the documents with their usual tools nor they can benefit from the flexibility of paper for indexing or annotating what is displayed on the screen. Having a separate medium for annotating (the notebook), which does not consume space on the screen is also another advantage, especially when using small laptops or mobile devices. Finally, another important requirement is the need for the students to link their own notes with the digital work they perform at home, a capability that is not supported by these systems (except StuPad, however with StuPad pupils can only attach a keyboard-typed text to a whole lecture, they cannot attach more sophisticated digital content such as web pages). Moreover, they cannot link this content precisely enough to link it to a specific sentence they may have written during the class.

2.2 Note Sharing and Annotation Tools

Miura *et al.* [11] present AirTransNote an interactive learning system that provides students with digital pens and PDAs. AirTransNote collects the handwritten drawings of the students and transmits them to the teacher's PC, so that teachers can closely monitor their students' work. A second version of AirTransNote [10] allows the teacher to replay the students' notes on a PC and to provide feedback on the PDAs of the students. These two works involved digital pens based on ultrasonic technology. When the student puts the pen down, ultrasonic waves are generated and provide the tip position relative to a sensor plugged at the top of the sheet of paper. However, the student has to specify when he starts to write on a new sheet of paper and cannot

modify previously written pages. The third version of AirTransnote exploited the Anoto technology [2] as a way to avoid these limitations. The Anoto technology uses a small camera embedded in a ballpoint pen to read a dot-pattern printed on paper in order to locate the pen's position. Although Miura *et al.* investigated the various versions of AirTransNote during experimental lectures at a senior high school, their studies mainly focused on note sharing and real-time feedback for students during short tests in class.

Using CoScribe [16], the teacher starts the lesson by giving printouts of the slides to the students, who can then directly create handwritten annotation on the teacher's printouts using an Anoto pen. They also can structure and tag their annotations for later retrieval. Finally, they can collaborate with other students by sharing their annotations. However, contrary to the system we propose, CoScribe uses the printouts as a central media and does not provide a way to associate the teacher's material with the notes in the student's notebooks.

2.3 Augmented Notebooks

The Audio Notebook [17] is a device combining a paper notebook with a graphical tablet and an audio recorder. The user can then listen to what was recorded when a specific note was written just by tapping on it. The Livescribe digital pens [9] extend this idea by including the audio recorder within an Anoto digital pen, making it possible to get rid of cumbersome devices such as the graphical tablet. However, these systems are limited to audio recording and cannot link handwritten notes with other types of digital data.

Other studies generally based on the Anoto technology have been devoted to augmented notebooks. Brandl *et al.* designed NiCEBook [3], an augmented notebook that enhances natural note taking. NiCEBook provides tagging functionality and allows users to share their handwritten notes in a vector format via e-mail. However, this system is not intended to link digital documents with personal notes.

Yeh *et al.* developed a notebook for field biologists [23] that associates handwritten notes with GPS coordinates, photos they shot or samples they found in the field. West *et al.* designed a similar system for scrapbooking [20]. Their system allows combining handwritten notes with media documents such as photos, videos and sounds using explicit gestures. Finally, Tabard *et al.* proposed Prism [18], a hybrid notebook that aggregates streams of digital resources (documents, web pages, emails) with biologists' notebooks. Its long-term deployment showed that, among all the data streams that were aggregated, users tend to rely on one of them as their master reference (which was generally the paper notebook). All these studies focused on different contexts than the electronic classroom. While they share some similarities with our work (as they also rely on augmented notebooks), the requirements of our application domain are different. For instance, our system allows fine-grained correspondence between handwritten notes and specific locations in digital documents, a feature that was not needed in these previous systems. The ability to link and further access the digital events that occurred during the class, and to enrich and personalize this data later at home and in other contexts, constitutes important improvements over these previous technologies, in our high school teaching domain.

3 Motivations and Interviews

To better understand how French teachers and pupils currently use digital material in the classrooms, we visited a school located in inner Paris. Based on insights from interviews with the teachers, we developed three online questionnaires that we distributed to teachers and pupils.

3.1 Method

Interviews. We interviewed three teachers with pupils from middle school (11-15 years old) and high school (15-18 years old). Each teacher was interviewed separately for one hour. We focused on their use of paper and digital materials during ‘normal’ classes, practical classes, and outside of the classroom.

Questionnaires. We gathered information about the uses and needs of paper and new technologies by means of three questionnaires [25]: two for the teachers and one for the pupils. We asked one elementary and two high school teachers to give us feedback on preliminary versions of the teachers’ questionnaire. This helped us to rephrase some questions so that they would better match the learning practices. We asked teachers of several schools to complete the questionnaire online. Eighteen teachers (15 female, 3 male) answered the questions (5 in elementary school, 8 in middle school and 5 in high school). The elementary school teachers each taught multiple topics. The other teachers either taught Mathematics (5), Literature (3), English (2), History and Geography (1), Physics and Chemistry (1) or Economy and Management (1).

We then designed a second teachers’ questionnaire and a pupils’ questionnaire to confirm and complete the answers of the first questionnaire. Nine pupils answered the pupils’ questionnaire, 4 from middle school and 5 from high school. Twelve teachers (9 female, 3 male) answered the second teachers’ questionnaire (3 were teaching in elementary school, 5 in middle school and 4 in high school). As before, elementary school teachers each taught several topics while the other teachers either taught Mathematics (3), Foreign languages (3), Literature (2) or History and Geography (1).

3.2 Results

We identified whiteboards, books, and paper handouts, as the teachers’ main resources for knowledge keeping, sharing or editing. Pupils mainly relied on notes written on paper and handouts to record the lectures and learn their lessons.

Teachers used multimedia equipment such as computers and video projectors as a way to augment existing lectures with digital documents, but not systematically for all lessons. These digital materials are of different kinds, depending on the topic of the lesson: for instance audio materials in language classes, videos in history or biology classes, and interactive demonstrations in mathematics or biology classes. A major problem we identified is that these materials are not often available to pupils after the lecture. These modern digital materials cannot be printed out and distributed as hardcopy, they must be sent by email or posted on an online teaching portal. However this does not appear to be a widespread practice, except for teachers with technical skills, who generally preferred to put digital materials on their personal web sites.

Finally, we also found that elementary school teachers scarcely use digital materials compared to middle and high school teachers.

Next we summarize the results of the three questionnaires and analyze the most interesting points.

Teaching Materials. Not all of the teachers have easy access to multimedia equipment such as computers and video projectors. For instance 3 out of 12 reported difficulties in having access to a video projector connected to a computer as often as they would like, and 4 out of 12 (at the elementary school) do not have access at all (Fig. 1, left). Practical constraints, such as the scarcity of, and time needed for installing these devices, reduced their availability, although the teachers showed interested in using multimedia equipment.

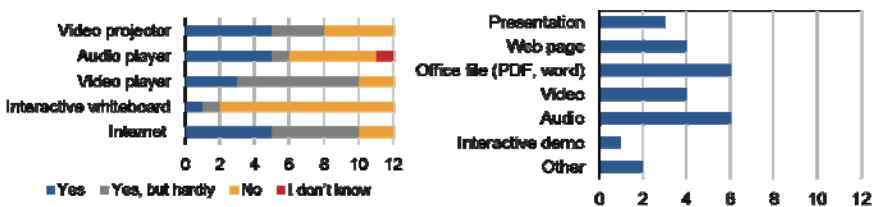


Fig. 1. Left: Number of teachers having access to these devices at school. Right: Number of teachers using a type of media during their lessons.

The analysis of the type of the media used during lessons (Fig. 1, right) shows that these resources are not necessarily in digital format: audio (used by 6 out of 12 of the teachers) and video (4 out of 12) could also be broadcasted in class using analog devices (e.g., VCR players). Non-digital media formats would be a barrier to sharing such documents with pupils. Other documents, such as pictures, exercises or tests in PDF formats are widely used but often printed on paper and distributed to pupils rather than projected during the class.

Four out of twelve of the teachers use web pages, which are used regardless of the lecture subject (Mathematics, History or Literature), this shows the potential of this media. Also, web pages are easy to share and flexible: teachers can give links or printouts, and they can also contain video or audio files.

Using digital materials in the classroom is a rather new practice. Even if relatively few teachers used them commonly, most of them thought this was going to increase in the future. As one of the high school teachers said: *“Last year, I renewed my lectures, and used many more slideshows and videos. The answer would have been very different two years ago: due to difficult access to the devices, I would not have bothered adapting or building my lectures around these resources.”*

Sharing Resources. We also observed that the sharing of materials was problematic. While half the teachers used video or audio resources, less than 10% sharing these resources after their lectures (Fig. 2, left). Teachers generally distributed their files as printouts that are distributed to pupils. However, this can affect the quality of the information. For example, audio transcriptions in foreign languages can be useful for

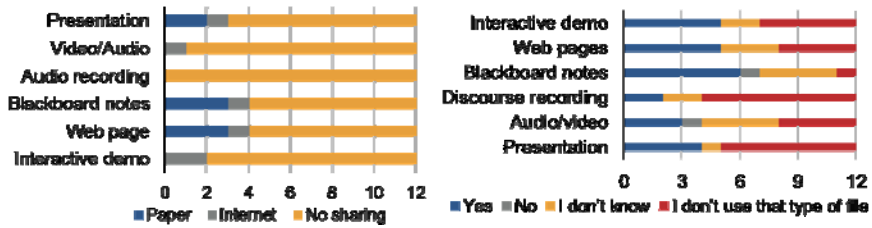


Fig. 2. Left: Do you share this type of file with your pupils? Right: Would you accept the creation of a history log of the files you use during your class?

practicing grammar and vocabulary, but is useless for practicing oral pronunciation. In addition, while some digital files like slideshows might be successfully conveyed on paper, other media like videos or interactive presentations would be difficult to convey through a printout if the material contains animation or other dynamic features.

Both pupils and teachers could take advantage of an easy sharing solution. Most teachers said they would accept using a system that would automatically transfer the digital files used during the lesson to the pupils (including a capture of the blackboard). But some teachers would only accept sharing files upon specific conditions such as the ability to control pupils' access and which files would be shared (typically, not the blackboard capture). As one teacher stated *"I'm in favor of transferring any type of digital data except the notes taken on the blackboard, to make sure of pupils take their own notes during the class."*

Capture in the Classroom. We further investigated whether teachers would agree to use a system that creates a log of the digital documents used during the lesson. As shown by figure 2 on the right, most teachers would accept that kind of system (given that they already used digital documents during their lectures) or do not know yet. In fact, less than 10% would refuse to use that kind of system for blackboard notes and audio/video files. Several teachers answered "I don't know", probably because of their concerns regarding the impact of sharing so much information with the pupils. As explained by one of the teachers, *"If everything is sent to pupils, they will not take notes anymore. Nevertheless, it can be useful for pupils to be able to access what happened in class to check a lesson from time to time or in case they missed a specific lecture"*. Hence, as noted above, providing the ability for teachers to control pupils' access to files is a key factor for acceptance.

Note Taking by Pupils. Pupils write notes on notebooks in all classes. Teachers progressively teach note taking to pupils, from the first years of middle school to high school. Initially the teachers write on the whiteboard, then progressively move to dictation, writing only keywords on the whiteboard. By the end of high school, pupils create their notes from the teachers' speech without the teachers having to provide written text. Yet, the teachers adapt the way of speaking from one class to another. As they dictate, teachers make sure that the pupils are still following or will slow down, moving from writing only keywords to the whole course on the whiteboard as necessary. Several teachers stressed that they wanted to ensure that their pupils take

notes. Hence, the principle of keeping the paper notebook as the central pupil's media, as we propose with U-Note, fits the teachers' recommendations.

Pupils not only take notes during lessons but also when working on computers in lab classes. For instance, during a visit to a high school, we observed pupils performing exercises on Open Office spreadsheets. They were asked to report results on paper printouts and explain how they solved the problem. Paper made it easy for the teacher not only to go through the pupils' work, but also to annotate the pupils' work and write comments and advice.

Without doubt, paper is still the most widely used media. As shown in figure 3 (left), pupils write in their notebooks on a daily basis and during most of the classes they attend. This observation was corroborated by the pupils' questionnaire answers. Two thirds of the teachers (12 out of 18) said that their pupils were writing on printouts everyday, and one third (6 out of 18) at least every week. This intensive use of handouts was not only explained by the good properties of paper (which is easy to use, to share, etc., as are notebooks) but also by the fact that all of the teachers we interviewed could easily access a photocopier (figure 1, left).

The amount of time spent in writing during the class is important (figure 3, right). According to teachers, most pupils spend more than 10 minutes writing in their notebooks during 55 minutes-long lessons (15 out of 18).

The use of paper was well summarized by one of the teachers: *“The notebook is the default medium. It is the only medium that lets us hope that pupils keep their materials from one class to the next one. I encourage pupils to write as much as possible on the handouts I give them so that they appropriate them, but it is somewhat difficult. They do not dare and feel reassured to write in their own notebook.”*

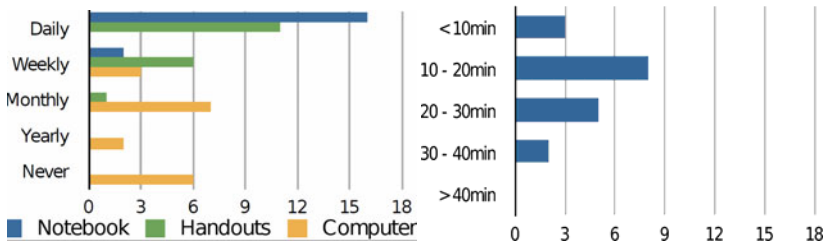


Fig. 3. Left: on which media do your pupils take their notes? Right: how much time do your pupils write during one hour of class?

Access to Notes. All pupils claimed to be reading their lessons at home¹. More interestingly, almost all of them (13 out of 15) declared reading their notes at school and about half (7 out of 15) declared reading their lessons while on public transportation. This underlines the importance of situations where the pupils cannot use their computer, and hence cannot access the teachers' materials.

¹ Even if the survey was anonymous, there may be some issues in trusting these numbers as the pupils answering the survey were likely the most dedicated ones.

3.3 Privacy Considerations

The remarks raised by the teachers pointed out some interesting concerns we initially overlooked, pertaining to who has access to the data. For example, many teachers felt that if the school administration could access their materials, the administration could also control how they work and what happened in the classroom. Teachers indicated that they wanted to keep the control of their own lectures. Some teachers also fear that these materials may be misused if given to the pupils. For example, some recordings could be posted on social networks to make fun of them (because of their accent, when they mumble or when they make mistakes). But one reluctant teacher said: *“If we can be assured that these recordings would only be used by pupils that want a new explanation of the lesson, I would definitely be for it. So, with strong guardrails, this could be interesting!”*

3.4 Implication for Design

This preliminary work indicated that paper is still the central medium for organizing information in the classroom. As we pointed out earlier, paper notes do not currently hold any type of digital information. In the following section, we present the system we proposed for augmenting paper notes with digital information. This system is based on the observations we made through the interviews and questionnaires presented above. On the teachers’ side, the system captures events during the class, and on the pupils’ side the events are linked to the pupils’ notebooks.

On the one hand teachers need:

- Devices for playing digital materials during their lectures.
- Digital materials (personal or academic data provided by institutions or editors).
- Systems for making these materials easy to store, share, and access. Additionally, these systems must provide access control so that teachers can specify what should be freely available to the pupils.

On the other hand pupils need a simple means for accessing the data in various situations (at home, in the library, in public transportation) and for associating it with their own notes. Hence:

- The notebook should remain the central media for pupils, but enhanced to make it possible to retrieve all useful information.
- The links between the pupil’s notes and the related digital media should be as specific and precise as possible in order to let pupils easily locate the information they are looking for.
- When pupils misunderstand some parts of the lesson they should be able to easily access the corresponding oral or written explanations, whenever possible.

4 Scenarios

Pupils use their notebook and the associated information in various locations (in the classroom, at home, etc.). Depending upon the situation, they may not have access to

the same devices and materials. We describe below several typical situations we identified as relevant.

4.1 In the Classroom

Ms. Green is giving a lecture on animals' breathing mechanisms in her life-science class. She introduces the lecture by raising questions regarding the breathing of different animals on earth, and in the water and air. Pupils interact with her and take notes on the introduction she dictates. After the introduction, she projects a video she had prepared earlier on her laptop.

The video presents breathing organs from three different animals: cows, crickets and salmon. Between each animal, she pauses the video and dictates to pupils what they just observed. While explaining the video she also draws diagrams on the whiteboard that pupils copy. Ms. Green then questions the pupils, so that they can progressively annotate the diagrams with arrows and labels.

4.2 At Home

A few days later Johnny, one of Ms. Green pupils, is doing his homework for the next class. While taking his notebook, he plugs his ANOTO pen to the computer to sync paper and digital notes.

The first exercise consists in identifying the different breathing organs of a frog. The case is complex as frogs use both lungs and skin to breathe. After checking his manual, he goes back to his paper notes and taps with his pen on the diagrams he drew in his notebook to be sure he did not forget any cases. The associated digital materials appear on his computer. While looking at the diagram, he notices a link to the video presented in the class, and loads it.

As Johnny can easily be distracted, he spots a link related to frog breathing in the comments and clicks on it. This link leads to a web tutorial that is helpful for his exercise. While reading it he uses the web capture tool (provided by his U-Note browser) to save captures of the most interesting parts to his digital notebook. These captures, especially the diagrams they contain, will be useful later when studying for the test. They may also be useful for his friend Frank, who often calls for help.

4.3 In a Mobile Situation

At the end of the month, Johnny has to prepare for his test. During a one-hour break, he goes to the library to review his lessons. As he suspects that the frog case or a similar one could be asked, he loads the link he saved a few weeks ago dealing with frog breathing and goes through it again. But, unfortunately, Johnny does not have enough time to finish reviewing before the next class. This is not really a problem as he will be able to continue in the bus, after the class, when going back home.

5 U-Note

U-Note was designed and developed to interact with notes and digital documents in all of the situations described above. U-Note comprises three tools. U-Teach is the

capture system used by the teacher. U-Move is the mobile client that can be used for browsing digital documents on a mobile phone. U-Study is the pupil software. It allows viewing and editing notes and the associated digital material on a PC. We describe these tools below through three tasks: capture, access, and annotation.

5.1 Capturing the Class

The classroom is the main capture location. The teacher provides information through speech, writing on the blackboard and digital documents. Meanwhile, pupils write their lesson in their notebook. Using a paper notebook is important for several reasons. Writing helps the student remember and understand [7] and using paper rather than computers prevents distraction [13]. Moreover, users' notes combined with audio recordings proved to be a powerful means of indexing meetings [21] (a situation similar to lectures). Furthermore, notes can serve as user-defined indexes for referencing events the user considers important [21].

The U-Teach module captures the information related to events occurring during the class while the pupils take notes. Teachers described in their interviews how they adapt the flow of their lessons to make sure that the pupils are still following. This ensures that the pupils' notes are synchronized with the teachers' discourse and the documents presented on the digital board. The U-Study module, described later, creates high granularity links between these events and the pupils' notes.

As noted earlier, teachers use multimedia files during their lectures, in particular in middle school and high school. The purpose of this part of the system is to record the context of the class. It is composed of several programs and plugins.

First we developed a PowerPoint extension that detects and records important events such as slide changes and the loading/unloading of presentation files. This plugin provides information regarding which slide of which presentation is shown at a given time. The same functionality is offered for web pages through a Firefox extension. As teachers often use audio or video recordings in their lectures, we also developed a dedicated multimedia player that logs actions such as load, unload, play and pause on these files. All these software components send events to a central server. This server generates a log file of the lecture that is accessible to the pupils' application. The current implementation does not check for access rights. While we did not focus on security issues, these could be resolved with a password system or certificates.

Finally we also capture the teacher's oral explanations using audio recording software that is running on the teacher's PC. The program allows the teacher to stop the recording, for instance if there is a disruption in the class. Additionally, the system can also take into account the events generated by an interactive whiteboard when such a device is available. The teacher's writing on the whiteboard is also made available to the pupils.

5.2 Accessing Digital Materials from the Notebook

Two tools are offered to pupils for accessing and enriching the information contained within the notes in their notebook: U-Study, a desktop application for working at

home, and U-Move, a web application that provides limited but still useful functionality in mobile situations.

U-Study. The U-Study module is a desktop application that displays a copy of the pupil’s handwritten notes and provides the digital documents used by the teacher during the lecture (Fig. 4). When reviewing a lesson, pupils may read any part they did not understand in class. By clicking on the corresponding notes in the notebook they can access the data related to this specific part of the course such as the oral recording at this specific moment, the slide, the web page or the video that was displayed at that time, and what the teacher was writing (depending on which media were used and captured during the class). Additionally, pupils can also open digital documents with their favorite applications, so that they can browse, and even edit and save them, more conveniently. We developed U-Study in Java with QT Jambi and used PaperToolkit [24] for retrieving the Anoto strokes from the digital pen.

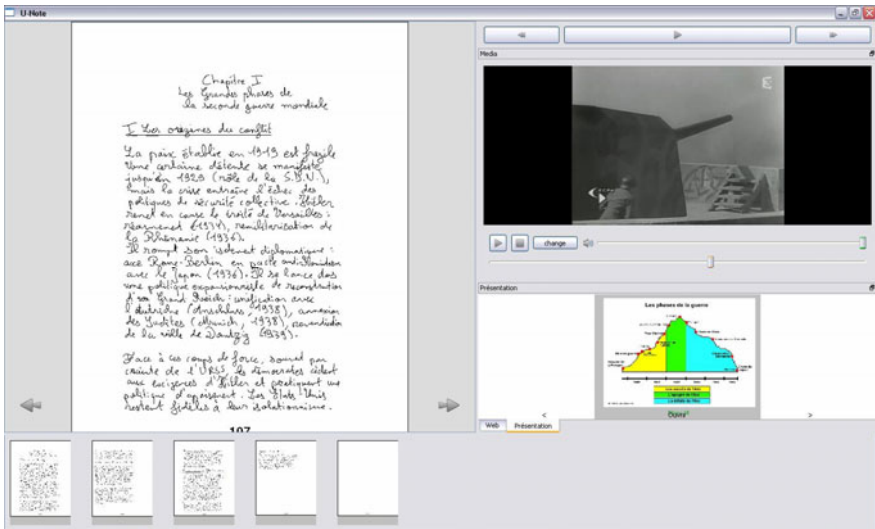


Fig. 4. Screenshot of the U-Study main view

The notebook view. The U-Study main window provides a view of the notebook (Fig. 5) that contains the strokes that were provided by the Anoto pen. The user can browse the pages of the notebook by clicking on two buttons. This view is mainly useful when the pupil’s notebook is not at hand, it can be hidden otherwise.

The miniature area. U-Study can also be used to explore the teacher’s documents while reading the notes. The “miniature area” can currently display four kinds of viewers (Fig. 6). The first viewer displays the miniature slides of a slideshow. It can be used for browsing the miniatures or for opening the original PowerPoint files. The second viewer allows the display of the web pages that were seen in class and to interact with dynamic content (hyperlinks, flash animations, etc.) when available. The third viewer is a multimedia player, which can play audio and video files. The fourth

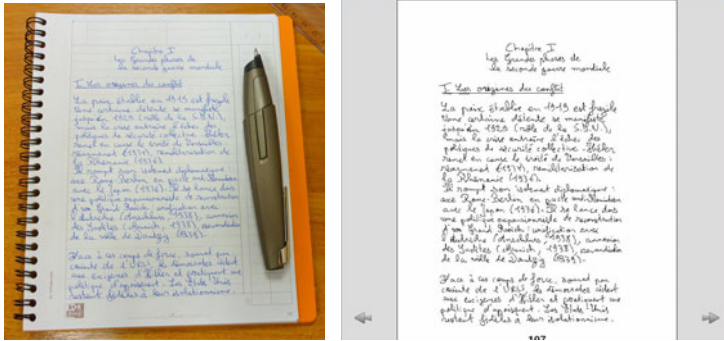


Fig. 5. Paper notebook and notebook view in U-Study. The pupil can use two buttons to browse the digital copy of his notebook.

viewer is an interactive whiteboard viewer that displays the teacher's drawings and writings if a whiteboard was used during the class. Any viewer can be displayed or hidden on demand. The user's favorite applications can also be used when preferred.



Fig. 6. Miniatures: PowerPoint, web pages and video

The thumbnails bar. The thumbnail bar provides a visual link between the notebook and the digital documents (Fig. 7). U-Study displays a thumbnail for each page of the paper notebook. When the pupil moves the mouse cursor, contextual tool-tips pop-up. The tool-tips contain a thumbnail of the specific parts of the documents that were displayed while the pupil was writing the page. Typically, each slide, video sequence, and web page has a corresponding thumbnail. When the pupil clicks on a thumbnail, the miniature of the document pops out in the appropriate miniature widget (as described above). In order to save screen space, U-Study displays six thumbnails simultaneously. Buttons located on the sides of the tooltip provide access to next or previous thumbnails if more than six documents are associated with the current page.

Replay. When the dynamic of the class is important, the pupil can “replay the class” from a given point. A red dot moves over the handwritten strokes on the notebook view to show what was written and when. The miniatures are updated in the corresponding views to show which materials (and which specific subparts of them) were shown in the class at that time. The pupil can interactively control the replay

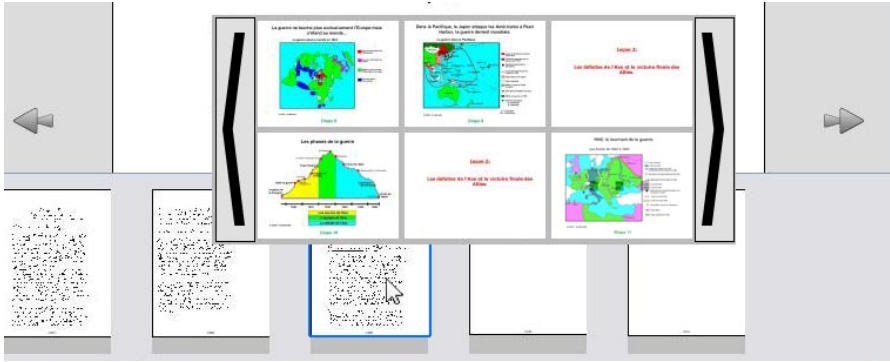


Fig. 7. Thumbnails bar

speed and pause or resume at any moment during the replay. Besides the red dot that continuously shows the temporal position in the notebook, any handwritten strokes written at a later time can optionally be grayed out to enhance the visual feedback (this is configurable by the user, as this feature may decrease readability).

Interaction with the notebook. The pupil can start a replay by tapping on their notebook with the Anoto pen connected in streaming mode with a PC. The system identifies the timestamp of the stroke specified by the pupil and the red cursor moves to the same position in the notebook view. The replay resumes at this time and the digital materials shown in class are opened in the miniature area. The fact that the notebook serves as a link between all the lecture materials makes this feature especially useful. The internal clock of the digital pen is synchronized to the clock of the PC each time the pupil plugs it to its computer. To ensure synchronization, the clocks in the pupils' and teachers' computers have to be synchronized via the network time protocol (NTP).

Unlike the previous systems discussed earlier [16,20,23], our system does not rely on explicit written marks (codes) since implicit correspondence between the strokes and the digital materials seemed more suited. First, pupils do not have to learn specific gestures and no error can occur because of the recognition algorithm. Second, as pupils do not control the flow of the lecture and since their attention is on the lecture, they may not have time or attention to dedicate to drawing explicit encoding marks. Finally, pupils can still write marks if they wish, using their own personal conventions. These marks will act as visual markers in their notebook (e.g., for highlighting an important aspect or for indicating a comprehension problem). Thanks to temporal associations, clicking on these marks will provide access to what the students expect. Hence, in most cases, there is no need for the system to understand the semantics of the user's marks and this would bring undesirable constraints such as forcing the pupil to use a predefined vocabulary of gestures.

U-Move. U-Move is a web application for mobile devices. We identified two main situations where a mobile application is useful. The first one is a fully mobile situation, as when the pupil is using public transportation, where the pupil wants to

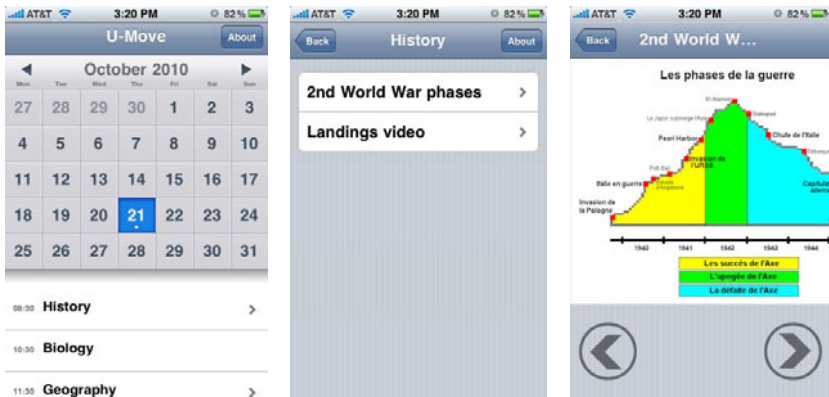


Fig. 8. Screenshots of U-Move - The U-Move calendar (left); Digital documents associated to a specific lesson (middle); A specific slide displayed on the mobile device (right)

look at the documents related to the lessons but does not have access to a PC, and manipulating his or her notebook may be somewhat cumbersome. The second situation is a less mobile situation, for example, when working in a library, in which the pupil does not have access to their own PC, but still wants to look at the lecture documents.

U-Move allows the pupil to browse the lectures documents. It consists of a calendar, which is synchronized with the pupil's schedule (Fig. 8, left). When the pupil taps on a day, the application displays the corresponding lectures. By tapping on a specific lecture, a list of the documents the teacher used that day is provided (Fig. 8, middle). These can then be opened by selecting them from the list; in which case the application downloads them from the central server and displays them (Fig. 8, right). The U-Move application has been developed as a Javascript web application based on the JQuery library [6] and the JQTouch plug-in [5]. We chose to develop this tool as a web application because this solution only requires mobile web access and can work on a variety of mobile devices, regardless of their operating system.

5.3 Extending the Lectures through the Notebook

While doing homework or studying lessons, the pupil will sometimes need to search for additional information on the web or other pedagogical resources. When useful information is found, these can be kept and paste into the notebook to make a link between the new document and the lesson.

Adding Digital Extracts to the Notebook. We developed a tool that allows adding pieces of documents in the digital notebook (Fig. 9). First, we developed a Firefox extension that allows one to capture a web page excerpt. This excerpt is a facsimile corresponding to the relevant subpart of the web page. The user interactively creates them by performing a drag selection (Fig.9, left). These excerpts remain attached to the original documents and can be refreshed and clicked as explained below. The U-Study module retrieves the document extracts sent by the capture tools through a

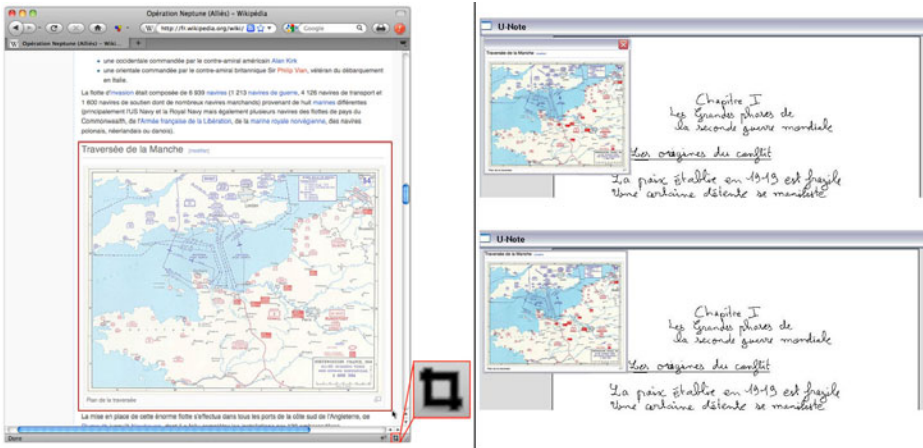


Fig. 9. Left: a pupil selects a region of a webpage that they want to stick to their notebook. The bottom right button is clicked, causing a bounding box to appear on the page. Right: the selected part can then be stuck to a page of the notebook.

socket. These excerpts appear as post-it windows (Fig. 9, right). An interesting feature is that they remain active so that the pupil can still click on the links and, for instance, view embedded files such as videos. A post-it can then be “stuck” to a given page of the notebook. Once stuck, the post-it is not active and cannot be resized to prevent unwanted modifications. It can be reactivated at will by unsticking it. The original related page can be opened in a web browser, so that these post-its are essentially bookmarks in the digital notebook.

Adding Physical Excerpts to the Notebook. We developed a tool that allows the pupil to print a physical interactive preview of any document opened on their PC. The pupil navigates in the U-Study menu to select the desired document to print. U-Study prints this document on Anoto paper and stores the mapping between it and the Anoto coordinates. The pupil can then cut and paste any part of the paper version of the document back into the notebook. The digital version may be opened by tapping on the piece of paper with the digital pen.

6 Conclusion

With U-Note we focused on helping pupils access the digital materials presented during classes. Preliminary interviews and questionnaires showed that while paper is still widely used, teachers are also increasingly using multimedia content. Consequently, we proposed to augment the pupil’s notebook so that it can serve as a central medium for referencing and accessing digital information. The notebook provides a simple means of accessing digital media presented in the class, together with oral explanations and the writings on the whiteboard. Moreover, the pupils can also enrich it by creating links to their own digital documents. U-Note allows fine-grained mapping between the notes and digital media and makes it possible to access them in various situations.

Future work includes a longitudinal study with teachers and pupils. We also plan to enhance the capabilities of the system, mainly for making it possible to capture more types of digital media and to easily create digital extracts from these files at precisely defined spatial or temporal locations. Finally, security aspects and access rights are also a topic we would like to address in future versions of the system.

Acknowledgments. We gratefully acknowledge the financial support of the Cap Digital ENEIDE project that was funded by Région Ile-de-France and DGE. We thank the anonymous reviewers of this article for their relevant recommendations.

References

1. Abowd, G.D.: Classroom 2000: an experiment with the instrumentation of a living educational environment. *IBM Systems Journal*, 508–530 (1999)
2. Anoto technology, <http://www.anoto.com>
3. Brandl, P., Richter, C., Haller, M.: Nicebook: supporting natural note taking. In: *Proc. CHI 2010*, pp. 599–608. ACM, New York (2010)
4. Brotherton, J.A., Abowd, G.D.: Lessons learned from eclass: Assessing automated capture and access in the classroom. *ACM Trans. on Comp.-Hum. Interact.*, 121–155 (2004)
5. Jqtouch plugin, <http://www.jqtouch.com>
6. Jquery library, <http://www.jquery.com>
7. Kiewra, K.A.: Note taking and review: the research and its implications. *Journal of Instructional Science*, 233–249 (1987)
8. Kong, C.K., Muppala, J.K.: ReCap: a tool for automated capture and generation of synchronized audio, PowerPoint and digital ink presentation. In: *Proc. CATE 2007* (2007)
9. Livescribe digital pens, <http://www.livescribe.com>
10. Miura, M., Kunifuji, S., Sakamoto, Y.: Airtransnote: An instant note sharing and reproducing system to support students learning. In: *Proc. ICALT 2007*, pp. 175–179. IEEE Computer Society, Los Alamitos (2007)
11. Miura, M., Kunifuji, S., Shizuki, B., Tanaka, J.: Augmented classroom: A paper-centric approach for collaborative learning system. In: Murakami, H., Nakashima, H., Tokuda, H., Yasumura, M. (eds.) *UCS 2004. LNCS*, vol. 3598, pp. 104–116. Springer, Heidelberg (2005)
12. Mühlhäuser, M., Trompler, C.: Digital lecture halls keep teachers in the mood and learners in the loop. In: *E-Learn, Montreal, Canada*, pp. 714–721 (October 2002)
13. Oviatt, S., Arthur, A., Cohen, J.: Quiet interfaces that help students think. In: *Proc. UIST 2006*, pp. 191–200. ACM, New York (2006)
14. Rössling, G., Trompler, C., Mühlhäuser, M., Köbler, S., Wolf, S.: Enhancing classroom lectures with digital sliding blackboards. *SIGCSE Bull.*, 218–222 (2004)
15. Sellen, A.J., Harper, R.H.: *The Myth of the Paperless Office*. MIT Press, Cambridge (2003)
16. Steimle, J., Brdiczka, O., Mühlhäuser, M.: Coscribe: Using paper for collaborative annotations in lectures. In: *Proc. ICALT 2008*, pp. 306–310. IEEE Computer Society, Los Alamitos (2008)
17. Stifelman, L., Arons, B., Schmandt, C.: The audio notebook: paper and pen interaction with structured speech. In: *Proc. CHI 2001*, pp. 182–189. ACM, New York (2001)

18. Tabard, A., Mackay, W.E., Eastmond, E.: From individual to collaborative: the evolution of prism, a hybrid laboratory notebook. In: Proc. CSCW 2008, pp. 569–578. ACM, New York (2008)
19. Truong, K.N., Abowd, G.D., Brotherton, J.A.: Personalizing the capture of public experiences. In: Proc. UIST 1999, pp. 121–130. ACM, New York (1999)
20. West, D., Quigley, A., Kay, J.: Memento: a digital-physical scrapbook for memory sharing. *Personal Ubiquitous Computing*, 313–328 (2007)
21. Whittaker, S., Tucker, S., Swampillai, K., Laban, R.: Design and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing* 12 (2008)
22. Wilkerson, M., Griswold, W.G., Simon, B.: Ubiquitous presenter: increasing student access and control in a digital lecturing environment. In: SIGCSE 2005, pp. 116–120. ACM, New York (2005)
23. Yeh, R., Liao, C., Klemmer, S., Guimbretière, F., Lee, B., Kakaradov, B., Stamberger, J., Paepcke, A.: Butterflynet: a mobile capture and access system for field biology research. In: Proc. CHI 2006, pp. 571–580. ACM, New York (2006)
24. Yeh, R., Paepcke, A., Klemmer, S.: Iterative design and evaluation of an event architecture for pen-and-paper interfaces. In: Proc. UIST 2008, pp. 111–120. ACM, New York (2008)
25. Zip file containing questionnaires,
<http://www.malacria.fr/unote/questionnaires.zip>

Erratum: Design and Evaluation of Interaction Technology for Medical Team Meetings

Alex Olwal, Oscar Frykholm, Kristina Groth, and Jonas Moll

School of Computer Science and Communication
KTH (Royal Institute of Technology), Stockholm, Sweden
{alx, frykholm, kicki, jomol}@csc.kth.se

P. Campos et al. (Eds.): INTERACT 2011, Part I, LNCS 6946, pp. 505--522, 2011.
© IFIP International Federation for Information Processing 2011

DOI 10.1007/978-3-642-23774-4_51

By mistake the following errors were introduced in the paper:

- Figure 1 was moved to the top of p. 506. It should have been placed on the bottom of p. 505.
- Figure 6 was moved to the middle of p. 513. It should have been placed on the bottom of p. 513.
- The end period of the captions of Figures 4 and 5 was removed.
- The authors in reference 18 were not formatted correctly. The correct version is: "Ratib, O., McCoy, J.M., McGill, D.R., Li, M., Brown, A."

The original online version for this chapter can be found at
http://dx.doi.org/10.1007/978-3-642-23774-4_42

Author Index

- Abascal, Julio IV-572
Abdelnour-Nocera, José IV-683, IV-738
Afonso, Ana Paula IV-300
Aigner, Wolfgang IV-292
Aizenbud-Reshef, Netta III-242
Aizpurua, Amaia IV-572
Alcañiz, Mariano I-523, II-44, II-65,
IV-475
Aldea, A. IV-466
Alem, Leila IV-442
Aliakseyeu, Dzmitry III-19, IV-744
Allmendinger, Katrin II-622
Almeida, Virgílio III-280
Alonso-Calero, José M^a IV-503
Alrayes, Amal III-680
Al-shamaileh, Ons IV-620
Alsos, Ole Andreas IV-84
Alvarez, Xenxo IV-358
Amir, Eyal I-348
Anacleto, Junia Coutinho II-422
Anand, Rangachari I-233
André, Elisabeth III-409
Andrews, Christopher II-589
Antila, Ville I-396
Antle, Alissa N. II-194, II-605
Anttila, Emilia II-554
Aquino, Nathalie IV-540, IV-677
Arroyo, Ernesto II-454, III-37, IV-462
Asakawa, Chieko I-83
Ávila, César IV-475
Avouris, Nikolaos I-489, IV-616
Axelrod, Lesley II-73
Ayoola, Idowu I.B.I. I-49
Aziz, Rafae I-426
- Baars, Arthur IV-640
Bachl, Stefan III-373
Baghaei, Nilufar IV-736
Bailey, Brian P. I-181, I-259
Baillie, Lynne II-36
Bailly, Gilles II-248
Balbino, Fernando Cesar II-422
Baldassarri, Sandra IV-495, IV-600
Ballano, Sergio IV-600
- Bannai, Yuichi III-323
Baños, Rosa M. IV-475
Barbosa, Glívia A.R. III-280
Barger, Artem III-242
Barreto, Mary I-129, IV-195
Bartram, Lyn II-52
Bastos, Pedro IV-358
Basu, Sumit III-187
Bautista, Susana I-57
Baxter, Gordon IV-736
Beale, Russell II-359, II-438, IV-632
Bellur, Saraswathi IV-487
Benedito, João I-65
Benevenuto, Fabrício III-280
Benford, Steve I-452
Benisch, Michael I-380
Bennett, Mike I-591
Benovoy, Mitchel I-152
Bérard, François II-107
Bergstrom, Tony I-627
Bernhaupt, Regina IV-716, IV-718
Besacier, Guillaume IV-366
Bevan, Nigel IV-652, IV-704, IV-732
Bevans, Allen II-194
Bezerianos, Anastasia IV-274
Bharshankar, Neha II-315
Bidwell, Nicola J. II-297, IV-738
Birnholz, Jeremy I-295
Blandford, Ann IV-728
Blat, Josep III-37, IV-462
Bogdan, Matthew I-108
Boll, Susanne II-640, III-134, IV-564
Booth, Kellogg S. III-338, IV-681
Bordbar, Behzad II-359
Bordegoni, Monica II-186
Bortolaso, Christophe II-341
Botelho, Fernanda IV-644
Botella, Cristina I-523, IV-475
Bowers, Christopher P. II-438
Boyd, Colin II-524
Bradel, Lauren II-589
Brandl, Peter II-1
Braune, Annerose III-610
Bravo, Crescencio IV-454

- Bravo-Lillo, Cristian IV-18
 Breiner, Kai IV-540, IV-724
 Brel, Christian IV-588
 Brereton, Margot II-524
 Brewer, Robin III-216
 Brewster, Stephen II-572
 Broll, Gregor IV-204
 Browatzki, Björn I-412
 Brunetti, Josep Maria IV-410
 Bruun, Anders IV-374
 Brynskov, Martin IV-685
 Buchholz, Gregor IV-511
 Buchner, Roland II-230
 Budzinski, Jochen III-477
 Bueno, André II-422
 Bullock-Rest, Natasha E. IV-689
 Burkhardt, Jean-Marie I-523, III-555
 Burri Gram-Hansen, Sandra IV-628
 Byrne, Will II-438
- Cabrera-Primo, David IV-580
 Cairns, Paul IV-309
 Calvary, Gaëlle IV-693
 Câmara, António I-1
 Camara, Souleymane IV-683
 Campos, Joana III-73
 Campos, Miguel IV-450
 Campos, Pedro IV-450, IV-730
 Cano-García, Josefa IV-503
 Carballar-Falcón, José Antonio III-265
 Cardet, Xavier IV-548
 Carmo, Maria Beatriz IV-300
 Carpendale, Sheelagh III-306
 Carriço, Luís IV-499, IV-734
 Carrión, Inma IV-36
 Carta, Tonio IV-349
 Caruso, Giandomenico II-186
 Casacuberta, Judit IV-527
 Casiez, Géry II-89
 Cassell, Jackie II-73
 Cearreta, Idoia IV-479, IV-572
 Cerezo, Eva IV-495, IV-600
 Chambel, Teresa I-152
 Chang, Ting-Ray IV-66
 Chapuis, Olivier I-215
 Chen, Fang II-178, IV-568
 Chen, Li III-234
 Chen, Monchu III-537
 Chen, Xiantao IV-491
 Chen, Yunan I-541, III-250
- Chisik, Yoram IV-100, IV-689
 Choi, Eric II-178
 Choong, Yee-Yin IV-1
 Chorianopoulos, Konstantinos I-199
 Christiansen, Lars Holm II-675
 Chynał, Piotr I-356
 Clemente, Miriam IV-475
 Clemmensen, Torkil IV-730, IV-738
 Clowney, Edrick I-108
 Cockburn, Andy I-331
 Collins, Patricia IV-596
 Coma, Inmaculada IV-483
 Comai, Sara IV-648
 Conde-Gonzalez, Miguel IV-604
 Condori-Fernández, Nelly IV-640
 Constantine, Larry III-537, IV-696,
 IV-706
 Conversy, Stéphane IV-531
 Cooperstock, Jeremy R. II-107
 Cordeil, Maxime IV-531
 Correia, Nuno I-609
 Couix, Stanislas III-555
 Coutrix, Céline III-54
 Cowan, Benjamin R. II-438
 Cox, Anna IV-178
 Coyette, Adrien IV-740
 Cranor, Lorrie Faith III-216, IV-18
 Creed, Chris II-438
 Cremonesi, Paolo III-152
 Curio, Cristóbal I-412
 Czerwinski, Mary I-2
- da Graça Campos Pimentel, Maria
 III-356
 Dalsgaard, Peter II-212
 Dang, Chi Tai III-409
 Darzentas, Jenny IV-720
 Davis, Matthew I-108
 Dayton, Will IV-596
 De Angeli, Antonella IV-620
 de Castro, Isabel Fernández IV-470
 Decker, Stefan I-388, IV-248
 Deller, Matthias II-289
 De Luca, Alexander IV-44
 de Magalhaes, Vanessa Maia Aguiar
 II-422
 Demarmels, Mischa II-622
 de Mattos Fortes, Renata Pontin
 III-356
 Dery-Pinna, Anne-Marie IV-588

- de Santana, Vagner Figuerêdo IV-349
 Detweiler, Christian IV-746
 Deutsch, Stephanie I-404
 de Wasseige, Olivier IV-693
 Dhillon, Beant II-392, IV-360
 Diamond Sharma, H. II-315
 Dias, A. Eduardo III-521
 Díaz, Marta IV-527
 Díaz, Paloma IV-726
 Dindler, Christian II-212
 Dittmar, Anke III-571
 Doering, Tanja II-605
 Domik, Gitta IV-722
 Dow, Lisa IV-736
 Downs, Julie IV-18
 Drexler, Felix IV-292
 Drucker, Steven M. III-187
 Du, Honglu I-199
 Du, Jia III-19
 Duarte, Luís IV-499
 Dubinsky, Yael III-242
 Dubois, Emmanuel II-341
- Ebert, Achim II-289, IV-722
 Edelmann, Jörg III-427
 Eggen, Berry II-263
 Elias, Micheline IV-274
 Elmqvist, Niklas III-391
 Endert, Alex II-589
 Engel, David I-412
 Epps, Julien IV-568
 Eskildsen, Søren II-297
 Essen, Harm van IV-744
 Evers, Vanessa IV-738
- Fairlie, Fiona II-36
 Faure, David IV-693, IV-740
 Fels, Sidney II-141, II-422, IV-691
 Fernández, Marcos IV-483
 Fernández-Alemán, Jose L. IV-36
 Fernquist, Jennifer III-338
 Ferre, Xavier IV-652
 Ferreira, Alfredo IV-523
 Ferreira, Denzil I-380
 Fetter, Mirko III-435, III-503
 Figueroa-Martinez, Jose IV-665
 Finnberg, Sanna III-169
 Firmenich, Sergio IV-340
 Fischer, Patrick IV-438
 Fisher, Danyel III-187
- Fitzpatrick, Geraldine II-73
 Flood, D. IV-466
 Forbrig, Peter III-571, IV-511, IV-718
 Förster, Florian II-230, IV-144
 Frederiksen, Nikolaj Yde II-675
 Frische, Florian IV-240
 Frykholm, Oscar I-505, E1
 Furtado, Elizabeth IV-679
- Gal, Eynat II-123
 Gallardo, Jesús IV-454
 Gallego, Fernando IV-454
 Gamecho, Borja IV-572
 Ganju, Aakash II-315
 Garay, Nestor IV-479, IV-572
 García, Arturo S. III-1, IV-612
 García, Roberto IV-410
 García-Herranz, Manuel I-576, II-479
 Garcia-Peñalvo, Francisco IV-604
 Gardiner, Steven I-426
 Garzotto, Franca III-152
 Gatti, Elia II-186
 Gault, Paul IV-118
 Gaunt, Kevin II-533
 Gázquez-Abad, Juan Carlos III-265
 Gerjets, Peter III-427
 Gerken, Jens II-622
 Gershon, Nahum IV-722
 Gervás, Pablo I-57
 Geyer, Florian III-477
 Ghani, Sohaib III-391
 Ghelawat, Sunil II-524
 Giannakos, Michail N. I-199
 Gil Iranzo, Rosa M. IV-519
 Gimeno, Jesús IV-483, IV-576
 Ginn, Tim III-117
 Giusti, Leonardo II-123
 Głomb, Przemysław IV-170
 Gomez-Aguilar, Diego IV-604
 Gonçalves, Daniel I-65, I-100, IV-644, IV-734
 Gonçalves, Glauber III-280
 Gonçalves, Jorge III-204
 Gonçalves, Tiago IV-300
 Goncu, Cagatay I-30
 González, María IV-535, IV-669
 González, Pascual III-1, IV-612, IV-665
 González-Calleros, Juan IV-740
 González-Deleito, Nicolás IV-636
 González-González, Carina S. IV-580

- Gonzalez Nieto, Juan II-524
 González Sánchez, José L. IV-519
 Goulati, Areti IV-360
 Graham, Nicholas II-341
 Granollers, Toni IV-418, IV-548, IV-656
 Grechenig, Thomas III-373
 Green, Lesley II-430
 Greenberg, Saul I-3, III-461
 Grisoni, Laurent II-89
 Gross, Sabine II-1
 Gross, Tom III-435, III-503
 Groth, Kristina I-505, E1
 Guerreiro, Tiago I-65, I-100, IV-644
 Guiffo, Joseph I-108
 Gunnarsson, Danial I-108
 Gürkök, Hayrettin I-115
 Gutierrez, Melvin IV-580
 Gutiérrez Vela, Francisco L. IV-519,
 IV-665
 Gutwin, Carl I-295, I-331
 Guy, Ido III-242
- Hackenberg, Linn IV-378
 Hajri, Abir Al II-141
 Häkkinen, Jonna II-333
 Hakvoort, Gido I-115
 Haller, Michael II-1
 Halskov, Kim II-212
 Hanks, Dustin I-108
 Harada, Susumu I-11
 Harris, Andrew I-627
 Harris, John I-108, III-45
 Harrison, R. IV-466
 Haya, Pablo A. I-576, II-479
 Heilig, Mathias II-622
 Heimgärtner, Rüdiger IV-738
 Held, Theo IV-438
 Helfenstein, Lukas E. IV-323
 Helmes, John II-376
 Hendley, Robert J. II-438
 Hennig, Stefan III-610
 Henwood, Flis II-73
 Henze, Niels III-134, IV-564
 Herdtweck, Christian I-412
 Hervás, Raquel I-57
 Heuten, Wilko II-640
 Hirose, Michitaka I-83
 Hiyama, Atsushi I-83
 Hobson, Stacy F. I-233
 Höchtl, Anita III-477
- Hoggan, Eve II-554
 Hoinkis, Monika I-470
 Holleis, Paul IV-204
 Holt, Behnjay I-108
 Holzinger, Andreas II-162
 Hooper, Clare J. IV-698
 Hourcade, Juan Pablo IV-689
 Hoven, Jeroen v.d. IV-746
 Huang, Ko-Hsun III-537
 Huang, Weidong IV-442
 Huber, Stephan II-622, IV-584
 Hucke, Maxi III-435
 Hupont, Isabelle IV-600
 Hurter, Christophe IV-531
 Hussein, Tim IV-726
 Hussmann, Heinrich IV-724
 Hutchings, Duke II-589
 Huuskonen, Pertti IV-592
 Huuskonen, Salla IV-152
 Hyatt, Alex I-248
- Ifukube, Tohru I-83
 Iglesias, Ana IV-535, IV-669
 Iivari, Netta III-288
 Ilich, Michael II-141
 Imhof, Birgit III-427
 Inami, Masahiko II-1
 Inkpen, Kori I-162, I-199
- Javed, Waqas III-391
 Jensen, Brit Susan II-675
 Jervis, Matthew III-100
 Jetter, Hans-Christian IV-584
 Jia, Haiyan IV-487
 Jiang, Dan-ning I-207
 Johansen, Silje I-470
 Johns, Paul I-199
 Johnson, Graham IV-118
 Jones, Allison I-470
 Jorge, Joaquim I-65, I-100
 Jorge, Joaquim A. III-461, IV-450
 Joshi, Anirudha I-313, II-315
 Jota, Ricardo III-461
 Jouffrais, Christophe IV-624
 Jurmu, Marko II-487
- Kaaresoja, Topi II-554
 Kaasinen, Eija IV-66
 Kaindl, Hermann IV-708, IV-712
 Kammerer, Yvonne III-427

- Kammoun, Slim IV-624
 Kappel, Karin III-373
 Kaptelinin, Victor I-444
 Kapuire, Gereon Koch II-297
 Karahalios, Karrie I-627
 Karapanos, Evangelos I-380, IV-195,
 IV-560
 Karppinen, Kaarina IV-446
 Karukka, Minna IV-592
 Katre, Dinesh IV-730
 Kauko, Jarmo II-333
 Kelley, Patrick Gage III-216
 Khaled, Rilla II-405
 Kim, Ki Joon II-281
 Kim, KyungTae III-391
 Kim, Taemie I-162
 Kimani, Stephen IV-736
 Kitchin, Mark I-108
 Kleindienst, Jan II-81
 Knoll, Avi IV-568
 Knolmayer, Gerhard F. IV-323
 Kobayashi, Masatomo I-83
 Kocielnik, Rafal II-392
 Komanduri, Saranga IV-18
 Komlodi, Anita II-471
 Koskela, Kaisa IV-446
 Kostakos, Vassilis I-380, II-487, III-204,
 IV-560
 Kow, Yong Ming III-250
 Kremer-Davidson, Shiri III-242
 Kristensen, Christian Haag III-662
 Kruger, Fabio II-487
 Kryski, Eric III-91
 Kuber, Ravi II-541, IV-458
 Kukka, Hannu II-487
 Kumarasamy, N. II-315
 Kun, Andrew L. IV-742
 Kurosu, Masaaki IV-738
- Ladeira, Ilda II-430
 Lai, Jannie IV-687
 Lai, Jennifer IV-256
 Laing, Angus I-362
 Landay, James A. I-11
 Langdridge, Darren I-362
 Lapidés, Paul III-45
 Larusdottir, Marta Kristin IV-430
 Law, Effie IV-714
 Lawson, J-Y. Lionel III-1
 Lazar, Jonathan I-108
- Lecolinet, Eric II-248
 Lecolinet, Éric I-643
 Lee, Bongshin I-162
 Lee, Juhnyoung I-233
 Lehtiö, Anu I-497, IV-592
 Leino, Juha III-169
 Leonardi, Chiara III-485
 Lepage, Matthew I-295
 Li, Jane I-248
 Liccardi, Ilaria I-215
 Lin, Jialiu I-380
 Lindgren, Helena III-644
 Lisai, Mauro IV-608
 Liu, Ning IV-491
 Liu, Ying IV-491
 Llinás, Pablo I-576, II-479
 Llórens, Roberto II-44
 Lohmann, Steffen IV-726
 Longo, Luca IV-402
 López, Juan Miguel IV-548
 López-Jaquero, Víctor IV-665
 Losada, Begoña IV-470
 Lozano, Jose A. II-44
 Lu, Jie IV-256
 Lucero, Andrés IV-744
 Lüdtke, Andreas IV-240
 Lui, Alfred I-396
 Lutters, Wayne II-471
 Luyten, Kris III-610
 Lynch, Sean III-306
 Lyons, Michael IV-691
 Lyra, Olga IV-560
- Macaulay, Catriona IV-118
 Macé, Marc J.-M. IV-624
 Macek, Jan II-81
 Macías, José A. IV-515
 Maciel, Cristiano IV-679
 Mackay, Wendy I-215
 Macq, Benoit III-1
 Madrid, Jaisiel IV-527
 Magnusson, Camilla IV-446
 Mahmud, Abdullah Al I-49, IV-661
 Makri, Stephann IV-728
 Malacria, Sylvain I-643
 Mancilla-Caceres, Juan F. I-348
 Marco, Javier IV-495
 Marcus, Nadine IV-552
 Marín-Clavijo, Jesus IV-503
 Markopoulos, Panos IV-360

- Marquardt, Nicolai III-461
 Marriott, Kim I-30
 Marsden, Gary II-430
 Marshall, Joe I-452
 Martens, Jean-Bernard I-49
 Martin, C. IV-466
 Martinez, Bibiana IV-576
 Martínez, Diego III-1, IV-612
 Martínez, Jonatan IV-612
 Martínez, Paloma IV-535, IV-669
 Martínez-López, Francisco J. III-265
 Martinie, Célia III-589
 Masip, Lúcia IV-418, IV-548, IV-656
 Mason, Jon IV-744
 Masoodian, Masood III-100
 Masthoff, Judith IV-118
 Matter, Inu I-277
 Mattes, Lisa IV-438
 Maurer, Max-Emanuel IV-44
 Mayer, Yael III-216
 Mazza, Davide IV-648
 Mazzone, Emanuela IV-632
 McAuley, Derek I-452
 McCarthy, Kevin I-591
 McCay-Peet, Lori IV-398, IV-728
 McGee-Lennon, Marilyn II-572
 Meerbeek, Bernt IV-744
 Meixner, Gerrit IV-540, IV-724
 Mendes, Daniel IV-523
 Mesa-Gresa, Patricia II-44
 Meschtscherjakov, Alexander II-657,
 IV-675, IV-742
 Miksch, Silvia IV-292
 Miller, Gregor II-141
 Minocha, Shailey I-362
 Miñón, Raúl IV-572
 Mirisae, Seyed II-524
 Mirnig, Nicole II-230
 Mirza, Iram IV-687
 Miura, Takahiro I-83
 Moeckel, Caroline IV-406
 Moere, Andrew Vande I-470
 Molina, Ana Isabel IV-454
 Molina, José P. III-1, IV-612
 Molina, Martin III-636
 Moll, Jonas I-505, E1
 Montoro, Germán I-576
 Moore, John IV-683
 Moreno, Lourdes IV-535, IV-669
 Morse, Emile IV-1
 Motayne, Mark I-108
 Motti, Vivian Genaro IV-700
 Müller, Jörg I-559, II-248
 Müller-Tomfelde, Christian I-248
 Murer, Martin II-657
 Murphy, Emma II-541
 Nacenta, Miguel A. III-306
 Nakajima, Kentaro IV-507
 Naumann, Anja I-559
 Navarro, María Dolores II-44
 Nawaz, Ather IV-390
 Negro, Sara III-152
 Nicholas Graham, T.C. II-18
 Nicholson, Amanda II-73
 Nicolau, Hugo I-65, I-100
 Nigay, Laurence III-54
 Nii, Hideaki II-1
 Ning, Tongyan II-248
 Nissenbaum, Helen IV-746
 Nóbrega, Leonel III-537, IV-136
 Noé, Enrique II-44
 Noguchi, Daisuke III-323
 Nore, Ville IV-446
 North, Chris II-589
 Nunes, Nuno Jardim III-537, IV-136,
 IV-195
 Nussbaumer, Philipp I-277
 Nzokou, Roslin I-108
 Oakley, Ian I-129, IV-556
 Obbink, Michel I-115
 Obrenović, Željko IV-710
 Obrist, Marianna IV-144
 Occhialini, Valentina II-263
 Offermans, Serge IV-744
 Oh, Jeeyun IV-487
 O'Hara, Kenton II-376
 Ojala, Timo II-487
 Okada, Kenichi III-323
 Oladimeji, Patrick IV-178
 Olanda, Ricardo IV-576
 Oliva, Marta IV-418, IV-656
 Oliveira, Eva I-152
 Oliveira, João I-65, I-100
 Olivera, Fernando II-479
 Olwal, Alex I-505, E1
 O'Modhrain, Sile I-591
 Opozda, Sebastian IV-170
 Oppl, Stefan III-443

- Orehovački, Tihomir IV-382
 Oriola, Bernard IV-624
 Orngreen, Rikke IV-730
 Orvalho, Veronica IV-358
 Osswald, Sebastian II-657
 Othman, Mohd Kamal IV-92
 Oyugi, Cecilia IV-683
- Paiva, Ana III-73
 Palanque, Philippe III-589
 Paldanius, Mikko IV-592
 Paletta, Lucas I-404
 Pan, Shimei IV-256
 Pan, Ying-xin I-207
 Panach, José Ignacio IV-640, IV-677
 Papachristos, Eleftherios I-489, IV-616
 Papadopoulos, Alessandro Vittorio
 III-152
 Parodi, Enrique III-636
 Pascual, Afra IV-548
 Pastor, Óscar IV-640, IV-677
 Paternò, Fabio III-628, IV-349, IV-608
 Patrício, Lia IV-136
 Paul, Celeste Lyn II-471, IV-1
 Pejtersen, Annelise Mark IV-730
 Petrie, Helen IV-92, IV-309, IV-720
 Pfeil, Ulrike III-477
 Pielot, Martin II-640
 Pietrzak, Thomas I-643
 Pinho, Márcio Sarroglia III-662
 Pittarello, Fabio I-144
 Plass-Oude Bos, Danny I-115
 Pleuss, Andreas IV-724
 Poel, Mannes I-115
 Pohl, Margit IV-292
 Politis, Ioannis II-392, IV-360
 Pombinho, Paulo IV-300
 Pommeranz, Alina IV-746
 Poon, Christina I-259
 Poppinga, Benjamin II-640
 Power, Christopher IV-92, IV-309
 Power, Richard I-57
 Prates, Raquel O. III-280, IV-679
 Pschetz, Larissa IV-426
 Puikkonen, Arto I-497
 Pujari, Sanjay II-315
- Qian, Huimin IV-458
 Quintal, Filipe I-129
- Raczewska, Agata IV-360
 Radke, Kenneth II-524
 Rähkä, Kari-Jouko III-169
 Rait, Greta II-73
 Ranch, Alex II-675
 Rane, Mandar II-315
 Ravendran, Rajinesh IV-434
 Read, Janet C. IV-632, IV-689
 Realinho, Valentim III-521
 Rehm, Matthias II-297
 Reis, Sofia I-609
 Reitberger, Wolfgang IV-144
 Reiterer, Harald II-622, III-477, IV-584
 Ren, Xiangshi IV-222
 Renevier-Gonin, Philippe IV-588
 Rennick Egglestone, Stefan I-452
 Ressin, Malte IV-683
 Rey, Beatriz IV-475
 Reyes-Lecuona, Arcadio IV-503
 Reynolds, Bernardo III-204
 Ribeiro, Nuno I-152
 Richter, Christoph II-1
 Rieder, Rafael III-662
 Righi, Valeria IV-462
 Rind, Alexander IV-292
 Rivard, Kathryn I-426
 Riveill, Michel IV-588
 Roberts, Dave I-362
 Robertson, George I-162
 Rodgers, Johnny II-52
 Rodil, Kasper II-297
 Rodrigues, Eduarda Mendes IV-364
 Rodríguez, Aina IV-475
 Rodríguez, Alejandro IV-475
 Rohs, Michael I-559
 Rojas, Luis A. IV-515
 Romão, Teresa III-521
 Romaszewski, Michał IV-170
 Romeu, Ignacio IV-640
 Rosales, Andrea III-37
 Rossi, Gustavo IV-340
 Roto, Virpi IV-714
 Roy, Debjani II-315
 Ruiz, Natalie IV-568
 Rupérez, María José II-65
 Russell-Rose, Tony IV-702
 Rutten, Romain II-315
- Sabatucci, Luca III-485
 Sadeh, Norman I-380, III-216

- Saini, Privender IV-661
 Sainz, Fausto IV-527
 Sali, Shweta II-315
 Samp, Krystian I-388, IV-248
 Sanchez, Fernando M. IV-576
 Sánchez-Franco, Manuel J. III-265
 Santoro, Carmen IV-608
 Saple, D.G. II-315
 Sarda, N.L. I-313
 Sato, Nozomi IV-507
 Sauer, Stefan IV-724
 Scarr, Joey I-331
 Schärfe,, Henrik IV-628
 Scherer, Reinhold II-162
 Scherr, Maximilian IV-204
 Schlögl, Stephan IV-422
 Schmehl, Susanne I-404
 Schmidt, Albrecht II-333, IV-742
 Schmieder, Paul IV-386
 Schmitz, Felix M. II-533
 Schossleitner, Gerold II-1
 Schrammel, Johann I-404, IV-394
 Schrempf, Andreas II-1
 Schrepp, Martin IV-438
 Schultz, Johnathon I-295
 Scott, Michelle I-129
 Seaborn, Katie II-194
 Sears, Andrew IV-458
 Sedaghat, Leela I-108
 Sedlins, Mara I-162
 Segawa, Ryo III-323
 Seifert, Julian III-503
 Seissler, Marc IV-540
 Selker, Ted II-454, IV-596
 Seneler, Cagla IV-309
 Sharlin, Ehud III-45, III-91
 Sharmin, Moushumi I-181
 Shoemaker, Garth III-338
 Shresta, Sujjan IV-683
 Signorile, Robert IV-186
 Sigut-Saavedra, Jose IV-580
 Silva, Hugo IV-644
 Silva, Ismael S. III-280
 Silva, Marcos Alexandre Rose II-422
 Silva, Margarida Lucas da IV-644
 Silva, Paula Alexandra IV-364
 Silveira, Milene IV-679
 Simões, Carla IV-644
 Simpson, Becky II-73
 Sinclair, Mike I-162
 Sini, Viola IV-323
 Sirkin, David I-162
 Sisti, Christian III-628
 Skelton, Dawn II-36
 Skov, Mikael B. II-675
 Sleeper, Manya IV-18
 Smith, Andy IV-683
 Smith, Helen II-73
 Smyth, Barry I-591
 Smyth, Michael IV-685
 Sochan, Arkadiusz IV-170
 Solomon, Davidson II-315
 Soukoreff, R. William IV-222
 Sousa, Cátia IV-556
 Sousa, Mario Costa III-45
 Sousa, Tiago Boldt IV-364
 Spano, Lucio Davide IV-608
 Speed, Chris IV-685
 Speelpenning, Tess II-605
 Stach, Tadeusz II-18
 Stary, Chris III-443
 Stent, Amanda III-636
 Stern, Kathryn I-108
 Stockinger, Tobias IV-44
 Stolze, Markus II-533
 Straßer, Wolfgang III-427
 Stroulia, Eleni IV-681
 Suárez, César Ortea IV-495
 Subramanian, Sriram III-19
 Sugimoto, Maki II-1
 Sugimoto, Sayumi III-323
 Sultanum, Nicole III-45
 Sundar, S. Shyam II-281, IV-487
 Susi, Angelo III-485
 Sutcliffe, Alistair IV-620
 Sutcliffe, Alistair G. III-680
 Sutton, D. IV-466
 Svanæs, Dag IV-84
 Sweller, John IV-552
 Swerts, Marc II-392
 Szafir, Dan IV-186
 Szostak, Dalila II-392
 Szymański, Jerzy M. I-356
 Tabard, Aurélien I-643
 Tak, Susanne I-331
 Tanenbaum, Joshua II-194
 Tang, John I-162
 Tarrago, Roger IV-462
 Taylor, Alex II-376

- Teixeira, Jorge IV-136
 Tenreiro, Pedro IV-364
 Thackray, Liz II-73
 Theofanos, Mary IV-1
 Theron, Roberto IV-604
 Thimbleby, Harold IV-178
 Thiruravichandran, Nissanthan II-675
 Thoma, Volker IV-544
 Tikkanen, Ruut III-288
 Toch, Eran I-380
 Tomasic, Anthony I-426
 Tomitsch, Martin I-470, III-373
 Toms, Elaine G. IV-728
 Tourwé, Tom IV-636
 Toval, Ambrosio IV-36
 Tran, Minh Quang I-362
 Trefz, Elmar I-470
 Tretter, Matthew II-541
 Tscheligi, Manfred I-404, II-230, II-657,
 IV-144, IV-675, IV-742
 Tsioporkova, Elena IV-636
 Tsoi, Ho Keung III-234
 Turic, Thomas IV-292
 Turrin, Roberto III-152
- Ukpong, Nsemeke IV-661
 Urretavizcaya, Maite IV-470
 Uzor, Stephen II-36
- Väänänen-Vainio-Mattila, Kaisa
 IV-714
 Vakkari, Pertti IV-152
 Valderrama-Bahamondez, Elba
 del Carmen II-333
 Van Dam, Joris II-315
 Van den Bergh, Jan III-610, IV-724
 van den Hoven, Elise II-605
 Vanderdonckt, Jean III-1, IV-693,
 IV-700, IV-740
 van Essen, Harm II-263
 van Tonder, Bradley II-505
 Varoudis, Tasos IV-52
 Vatavu, Radu-Daniel II-89
 Vatrapu, Ravi IV-738
 Veer, Gerrit van der IV-722
 Venkatanathan, Jayant I-380, III-204
 Venolia, Gina I-162
 Vera, Lucía IV-483
 Vermeeren, Arnold IV-714
- Vertegaal, Roel III-117
 Vigo, Markel IV-734
 Vilar, Nicolas II-376
 Vilhelm Dinesen, Jens IV-628
 Virolainen, Antti I-497
 Vogel, Daniel II-89
 Vogt, Katherine II-589
 Vos, Tanja IV-640
- Wählen, Herje I-444
 Waite, M. IV-466
 Walker, Brendan I-452
 Walter, Robert II-248
 Wang, Guangyu II-107
 Wang, Sijie II-194
 Wang, Xiaojie IV-491
 Wang, Yang II-178, IV-568
 Warnock, David II-572
 Watanabe, Willian Massami III-356
 Weber, Gerhard IV-720
 Weiss, Astrid II-230, IV-675
 Weiss, Patrice T. II-123
 Wen, Jing III-250
 Wen, Zhen IV-256
 Wentz, Brian I-108
 Wesson, Janet II-505, IV-718
 White, Elliott P. IV-544
 Wiehr, Christian IV-540
 Wightman, Doug III-117
 Wiles, Alison IV-683
 Wilfinger, David II-657, IV-675, IV-742
 Williams, Sandra I-57
 Wiltner, Sylvia IV-292
 Winckler, Marco III-589, IV-340,
 IV-679, IV-718
 Winnberg, Patrik J. III-644
 Winnberg, Peter III-644
 Winschiers-Theophilus, Heike
 II-297, IV-738
 Wobbrock, Jacob O. I-11
 Wolf, Katrin I-559, IV-414
 Wolfe, Christopher II-341
 Wong, Anna IV-552
 Woodbury, Rob II-52
 Wrzesien, Maja I-523, II-65
 Wurhofer, Daniela IV-144
- Xu, Jie II-178, IV-568

- Yang, Jeaha I-233
Yao, Lin I-207
Yeo, Alvin IV-738
Yeung, Ching-Man Au I-215
Young, James III-45

Zancanaro, Massimo II-123, III-485
Zarjam, Pega IV-568
Zetterström, Erik IV-362
Zhang, Aiping IV-1

Zhao, Jian IV-222
Ziefle, Martina II-162
Ziegler, Jürgen IV-726
Zilouchian Moghaddam, Roshanak
I-259
Zimmerman, John I-426
Zlotowski, Jakub IV-370
Zöllner, Michael IV-584
Zwartkruis-Pelgrim, Elly III-19